אוניברסיטת
בר־אילן
Bar-Ilan University

Faculty of Engineering

Bio-Engineering Laboratory

Development and Optimization of a Machine-Learning Prediction Model for PDT and PTT Therapy in Solid Tumors

Hadas Dahan

Sarah Bitton

Professional Advisor: Eli Varon

Academic Advisor: Prof. Orit Shefi

October 2022

Table of content

Abstract

A major challenge in radiation oncology is to predict and optimize a clinical response on a personalized basis. New approaches are needed to deliver a customized light treatment with consistency, low side effects, and most importantly patients' outcome. Lately, Artificial Intelligence (AI) and particularly machine learning subsequent decisions are implemented in different areas of light therapy, such as target volume, geometric complexity, image segmentation, and model quality assurance. Yet today there are no clear guidelines on commissioning an optimization procedure to determine treatment parameters for Radiation Therapy (RT) particularly Photodynamic Therapy (PDT) and Photothermal Therapy (PTT).

To address this challenge, we developed a gold nanoparticle (AuNP) conjugated with a PDT agent, meso-tetrahydroxyphenylchlorin (mTHPC) photosensitizer as a nanotherapeutic complex to achieve a dual PDT/PTT therapy. The AuNP-mTHPC complex can serve as a dual drug delivery vehicle for improved single modality photodynamic or photothermal therapy. Also, the AuNP-mTHPC complex, is biocompatible, photostable, and induces a synergic anti-cancer effect. Following, we developed a ground-breaking algorithm that implements specific treatment parameters for PDT and PTT to achieve a personalized clinical outcome. Without loss of generalization, we focused on the laser power (mW/cm^2) and duration (min) parameters that lead to successful cell death. We examined common prediction models in machine learning to create a data set that contains training and test sets to demonstrate the suggested technology. Next, we determined the prediction model precision by cell death calculation following light treatment to predict each treatment efficiency independently and combined. Our goal is to create a model that will be able to determine the mortality rate of cancerous cells giving the wave intensity and the exposure time. We used different types of phototherapies, PDT and PTT, as well as different types of machine learning models, logistic regression, SVM (Support Vector Machine) and Decision Tree. Based on our results, we can now provide a machine learning based personal prediction model in a combined nanotechnology radiation treatment.

Acknowledgments

We would like to express our deep gratitude to Professor Orit Shefi, head of the Neuro-engineering and Regeneration laboratory and our academic supervisor for giving us the opportunity to research and providing us with guidance through the project.

In addition, we would like to thank our kind professional advisor Mr. Eli Varon for his guidance, patience and help throughout the entire project process.

We would also like to thank Dr. Gaddi Blumrosen for assisting us with data analysis in machine learning algorithms.

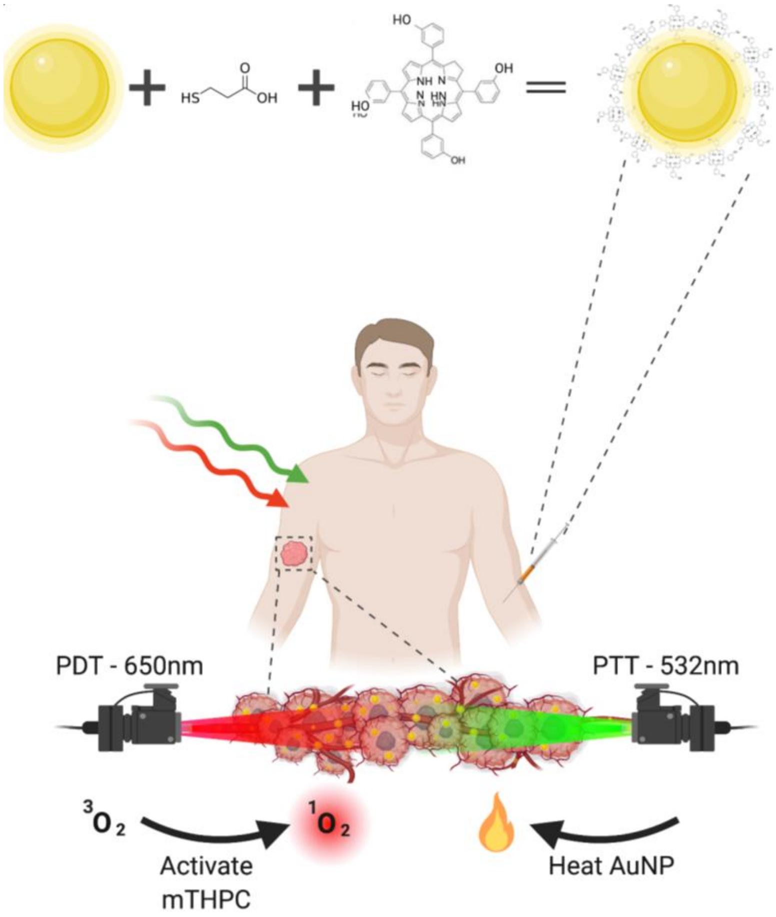
1. Introduction

Cancer is a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body. When abnormal or damaged cells grows and multiply they can form tumors. Tumors can be cancerous or benign. Cancerous tumors spread into nearby tissues and can travel to distant places in the body to form new tumors [1].

Cancer Radiation Therapy (RT) is one of the most used non-surgical interventions in tumor treatment. Traditionally, it is based on a high-level focused radiation elevated towards the solid tumor. The radiation kills cancer cells or slows their growth by damaging their DNA [2]. However, due to the proximity of surrounding normal tissue, movement of patient, and low oxygen supply radiation therapy remains non-specific and can lead to serious side effects. Hence, many researchers are developing targeted radiation methods that deliver a higher dose of radiation to the solid tumor while limiting damage to surrounding healthy tissues.

In recent years, the combination of nanotechnology with radiation have shown improved treatment efficacy and greater specificity. A novel class of radiation sensitizers comprises nanoparticles (NPs) which are highly efficient and selective platforms. These NP accumulate inside tumors due to the Enhanced Permeability and Retention effect [3]–[5]. Then a near-infrared light radiated to the desired area, can kill with enhanced selectivity the tumor.

Phototherapies involve the irradiation of target tissues with light. To further enhance selectivity and potency, numerous molecularly targeted photosensitizers and photoactive nanoparticles have been developed. Active targeting typically involves harnessing the affinity between a ligand and a cell surface receptor for improved accumulation in the targeted tissue. Targeting ligands including peptides, proteins, aptamers and small molecules have been explored for phototherapy [6]. We focused on two phototherapies and examining their effects: photodynamic therapy (PDT) and photo thermal therapy (PTT).



Project goal

We would like to predict which combination of our parameters: the laser intensity (mW/cm^2) and the exposure time (min) will cause higher mortality rate of cancerous cells and will yield better results as a treatment.

In order to achieve this goal, we compared between PDT and PTT treatment and between different machine learning algorithms to see which one is more precise and after knowing which algorithm have better result, we can enter to our algorithm a desirable set of parameters and know the mortality rate without needing to repeat the experiment for those parameters.

1.1 Photodynamic Therapy

Photodynamic therapy (PDT) is minimally invasive and clinically approved procedure for treating multiple types of cancer. In PDT we administer a photosensitizer (PS) followed by exposure to light. Upon the light absorption, the PS transforms from ground state to an excited single state. This excited state produces radical and reactive oxygen species [7], [8]. When the photosensitizer the light and oxygen react, it generates singlet oxygen. Singlet oxygen cause toxicity leading to cell death [9]. Currently there's wide range of photosensitizer with different properties. One of them is meso-tetrahydroxyphenylchlorin (mTHPC) which is a second-generation photodynamic sensitizer. mTHPC is a very effective photosensitizer that induce high phototoxicity in the cell at very low concentrations [10]. mTHPC is activated by illumination of light wavelength of $\lambda = 652$ nm in the red region and to achieve our dataset we used laser in that region with different wavelength and intensities.

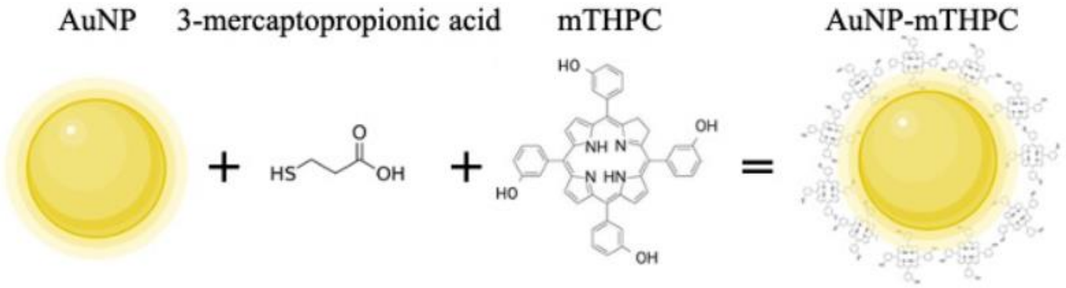
PDT as a treatment for cancer have advantages over other treatments. It has reduced long-term illness compared to other treatment and it does not compromise other treatment options.

1.2 Photothermal Therapy

Photothermal therapy (PTT) is minimally invasive treatment with low side effects. In PTT we inject photothermal agent, after it expose to light it converts photo energy into thermal energy [11]. When the photothermal agents, the nanoparticles, are exposed it cause synchronized oscillation of the nanoparticles conduction band electrons that results in heat. This causes temperature rise in the tumors that cause irreversible cellular damage and result in thermal ablation of the cell. The temperature rise depends on the nano particles concentration, the photothermal conversion efficiency and the dosage of light delivered to the cells. To maximize the probability of thermal ablation in the cells we use nano particles that designed to absorb near-infrared (NIR) wavelength, $\lambda \approx 650$-1064 nm, NIR can penetrate soft tissues making it possible to direct eradication of cancer cells in distant organs.[9]

2. Methods

2.1 Data collection



This interdisciplinary research project involves the combination of biotechnology, engineering, and AI. All these different studies combined, fundamentally alter the way radiation therapy is practiced. Firstly, the Shefi Lab developed a nanocomplex with clinical approved components and specific drug delivery characteristics. They decided to use spherical gold nanoparticle (AuNPs) that can be activated by a green laser (532 nm) because they are considered fine photothermal agents due to their surface plasmon resonance (SPR) effect, great light-to-heat conversion, simple surface functionalization for tumor targeting, and small diameters which allow solid tumor penetration.

Following, they chose Meso-tetrahydroxyphenylchlorin (mTHPC) as the conjugated photosensitizer because it has many advantages such as, strong phototoxicity, high triplet yield, high absorption (650 nm), and exceptional light penetration in tumor tissue. The temperature profile of the AuNP-mTHPC complex was measured using a radiometric thermal imaging camera. We grew SH-SY5Y neuroblastoma cells and incubated our cells with the AuNP-mTHPC complex. For the PDT therapy we apply laser beam at $\approx 650\ nm$ the red region to activate our drug (mTHPC). For the PTT therapy we apply laser beam at $\approx 532\ nm$ the green region to activate the gold nanoparticles.

We will collect a database which contains 3 different features: in inputs the laser intensity (mW/cm^2) and the exposure time (min) and from the experiments result (death %) will outputs the mortality rate of the cells.

We will use the collected database to create a model to determine the mortality rate of the cells giving only the wave intensity and the exposure time.

2.2 Prediction model

We are working with 3 features: The treatment (PDT/PTT), the exposure time (min) and the intensity of the laser (mW/cm^2).

With only the few data points we have at our disposal, around 100 data points, trying to directly predict the mortality rate is a very difficult problem for a model because it is a continuous target. That’s why we transform this target into a categorical one, creating mortality classes with intervals.

We tested two transformations of the mortality rate varying the thresholds enabling to assign a mortality rate to a class among 3 possible classes that are representing the mortality rate of the cancer. One transformation enables to create a balanced dataset in terms of classes repartition and the other leads to an unbalanced dataset. The exact rules enabling us to derive the two datasets are presented below:

$$\text{Unbalanced (According to Biology threshold): } \begin{matrix} U & > 30\% \\ A & 30 > x > 60 \\ D & < 60 \end{matrix}$$

$$\text{Balanced: } \begin{matrix} U & > 10\% \\ A & 10 > x > 36 \\ D & < 36 \end{matrix}$$

Where U is for Unaffected, A for Affected and D for Dead.

Unbalance data, or imbalance data, is data that have unbalance value across categorical variables, usually with biased towards on of the variable and very few in the other variables[12]. Most machine learning models are designed for a balanced problem, making them more likely to focus on learning the characteristics of the majority class neglecting learning from examples of minority class. This can cause the model to have a harder time predicting the minority class [13].

We have a few options to correct our dataset: Undersampling, Oversampling and Data Augmentation. In undersampling we discard several data points of the majority class, since our datasets is small, we can`t afford to discard data points. In oversampling we make duplicated of data points in the minority classes, in our data collection we repeated every experiment twice and duplicating the data point will create 4 data point for every experiment in the minority classes. Data augmentation works like oversampling, but the copied data points have small deviation [12].

We decided to use data augmentation as the solution for our unbalanced data and used the SMOTE algorithm. The algorithm creates synthetic observations using linear interpolations between the points in the minority classes.

We used 3 classifiers to create our models, Logistic regression, Support Vector Machine (SVM) and Decision Tree. Every classifier tries to predict for a given data point its class label, meaning U, A and D, if we were to have the experiment with the given parameters. The logistic regression classifier is trying to separate the data points with linear boundaries thanks to which we can derive the probabilities to belong to each class for each point. The Decision Tree is trying to find rules enabling to separate the data. We use the SVM classifier with 3 different types of kernels, with a linear kernel, which is also trying to separate the data points with linear boundaries but this time, it is searching for the linear boundaries that are maximizing the margins with the data points. With other type of kernels (gaussian or polynomial for instance), the data is represented in another space, and we try to separate the data points linearly in this new representation hoping that it is better suited than the original ones.

2.3 Validation methods

Usually, it is common to split the dataset into a training set (to make a model learn), a validation set (to tune some hyperparameters) and a test set (to see the ability of the model to generalize on unseen data). Here, we are working with very few data and hence we decided to only work with a train set and a test set. Nevertheless, the fact that our dataset is very small enables us to make a leave-one-out cross validation at practically no cost. That process consists of cutting our dataset into a train set and a test set as often as the number of observations, isolating each time one single observation for our test. It enables us to assess much more precisely the ability to generalize of our models.

In our data, we have redundancies since each experiment has been carried out twice to check that the assignment of the classes was not random. Hence, it is a bit problematic to use the data as it is, especially for the leave one out cross validation where we do not want the test points to be included in the train set. That is why we decided in accordance with our supervisors to add a random gaussian noise with very little standard deviation to simulate diversity in the data. Furthermore, we divided the dataset according to the treatment type PDT or PTTs.

To evaluate the performance of our models we need some parameters like the confusion matrix and the accuracy.

Confusion matrix

The confusion matrix is a matrix that enables to observe the quality of the predictions of a given classifier (traditionally with a threshold of 0.5 applied to the output probabilities) [14]. It contains the number of correct predictions per class as well as all the errors (predict class X instead of class Y). From that matrix, we can derive meaningful metrics such as the recall of each class (= proportion of correctly predicted X among all the X that we had to predict), the precision of each class (=proportion of correctly predicted X among all the X predicted) and the f1-score of each class (=harmonic mean of recall and precision).

In the diagonal, we have the number of correct predictions by class. The other cases correspond to errors. In our case, the i, j element of the matrix corresponds to the number of times the model predicted the class j instead of the class i.

Accuracy

The accuracy is the proportion of correct predictions given by our model on the test points. It’s the most intuitive metric used for classification but not the most relevant one in general, especially when we have an unbalanced dataset.

False positive/True positives

Our problem is not binary and hence we can’t talk about false positive and true positive. We have to exactly describe each error as being the error related to predicting class A instead of class B.

ROC curves and AUC

AUC is “the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative').” [15] It is a metric only relevant for binary classification problem.

Here, we can transform our model into 3 binary classification problems relative to each class where each time we create a model that aims to predict if a sample belongs to the corresponding class or not. It is called the One vs All strategy.

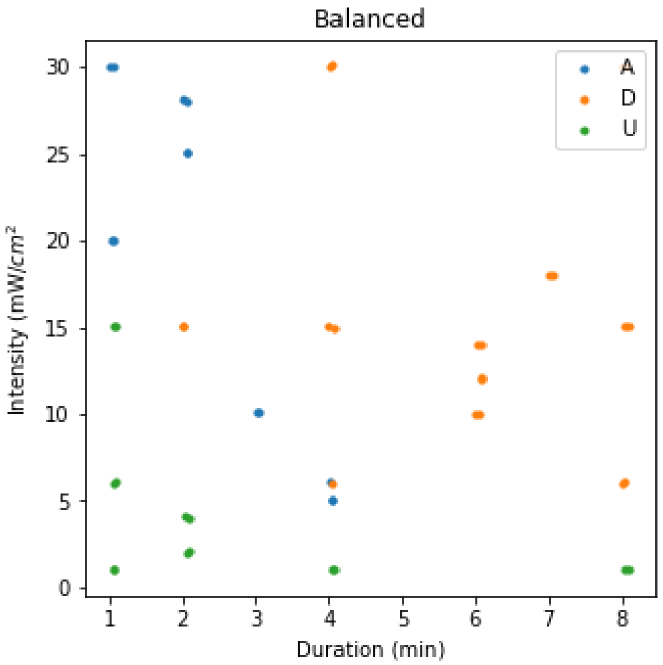
From the 3 binary classifiers created we can get 3 ROC Curves. Then, we can use these curves to derive:

- macro-AUC: Obtained doing the mean of the 3 ROC Curves
- micro-AUC: Obtained from the concatenation of all the outputs of the 3 classifiers.

These new metrics can then be used for comparisons with other models.

3. Results

3.1 Balanced Dataset for PDT Treatment

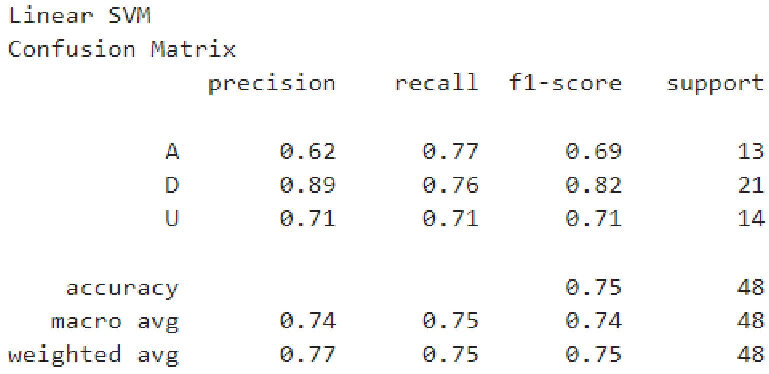


From this simple scatter plot, we can already see that the classes are more or less linearly separable. Hence, using a linear classifier such as the logistic regression and the LinearSVM should do the trick. Moreover, we can be also tempted to use boxes to separate the classes and so the decision tree is also a good choice.

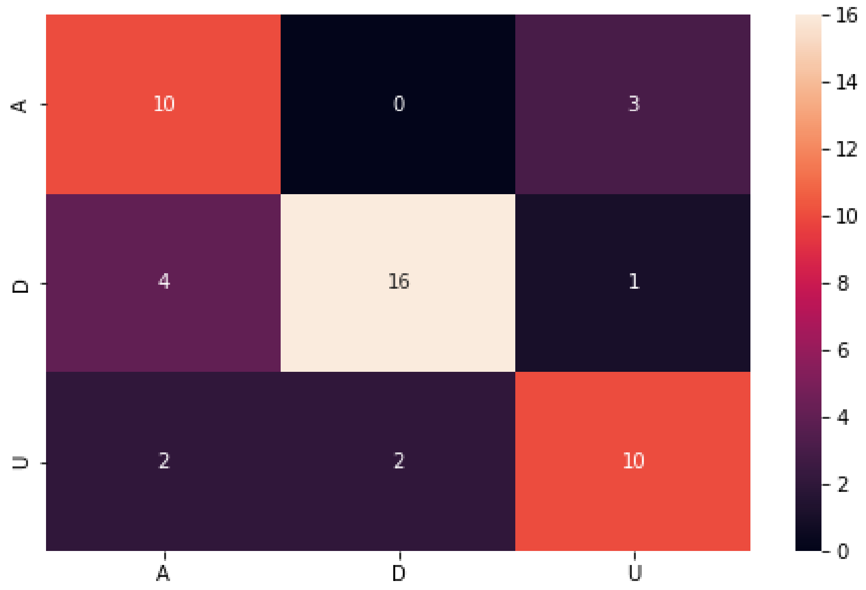
From this plot we can guess that the class A will not be easily guessed by a linear model considering that some points in this class are in the middle of the other points. So, we could expect a not so good performance for this class.

In the balanced PDT dataset we have 48 data points from them 13 data points have the Affected label, 21 data points have the Dead label and 14 data points have the Unaffected label. In the unbalanced dataset we have 8 data points have the Affected label, 14 data points have the Dead label and 26 data points have the Unaffected label. In the SMOTE dataset we have 78 data points with 26 data points in every class.

Confusion matrix for Linear SVM on Balanced dataset for PDT treatment:

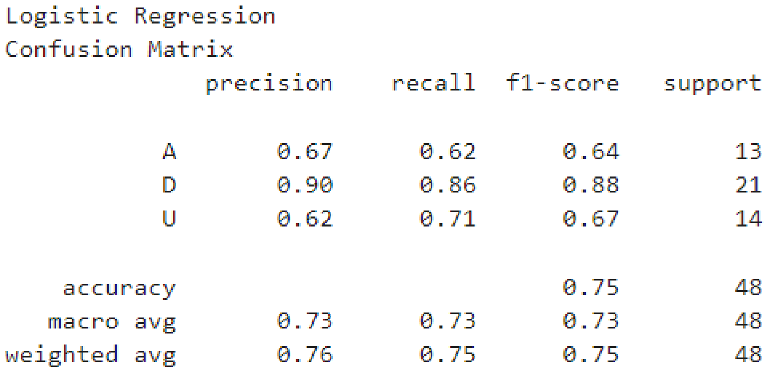


```
Linear SVM
Confusion Matrix
              precision    recall  f1-score   support

           A       0.62      0.77      0.69        13
           D       0.89      0.76      0.82        21
           U       0.71      0.71      0.71        14

    accuracy                           0.75        48
   macro avg       0.74      0.75      0.74        48
weighted avg       0.77      0.75      0.75        48
```

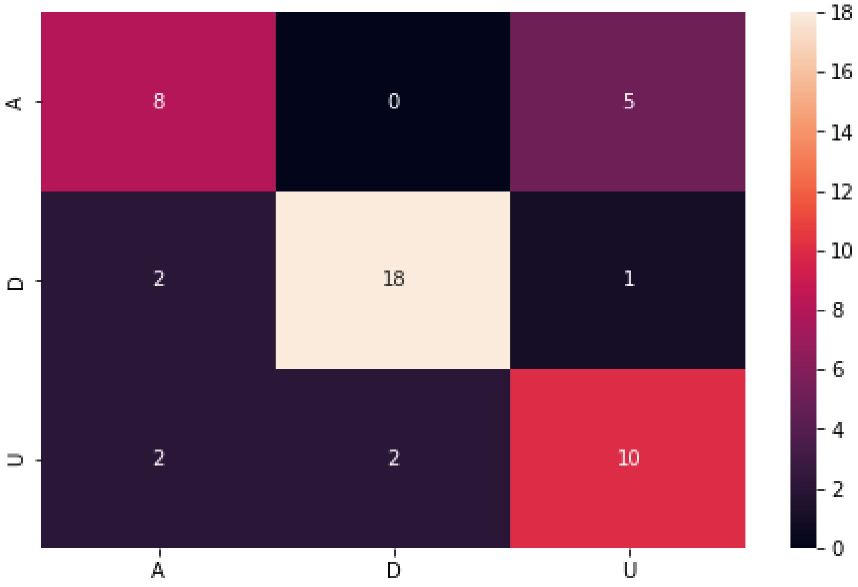


We can see from the diagonal in the confusion matrix that the Linear SVM have predicted 10 data points as Affected, 16 data points as Dead and 10 data points as Unaffected. From the first row we can see that the Linear SVM predicted 0 Affected data points as Dead and 3 Affected data points as Unaffected. From the second row we can see that the Linear SVM predicted 4 Dead data points as Affected and 1 Dead data points as Unaffected. From the third row we can see that the Linear SVM predicted 2 Unaffected data points as Affected and 2 Unaffected data points as Dead.

Confusion matrix for Logistic Regression on Balanced dataset for PDT treatment:

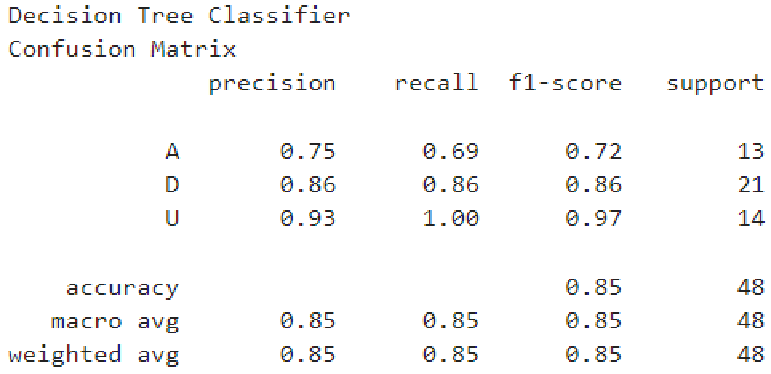


```
Logistic Regression
Confusion Matrix
              precision    recall  f1-score   support

           A       0.67      0.62      0.64        13
           D       0.90      0.86      0.88        21
           U       0.62      0.71      0.67        14

    accuracy                           0.75        48
   macro avg       0.73      0.73      0.73        48
weighted avg       0.76      0.75      0.75        48
```

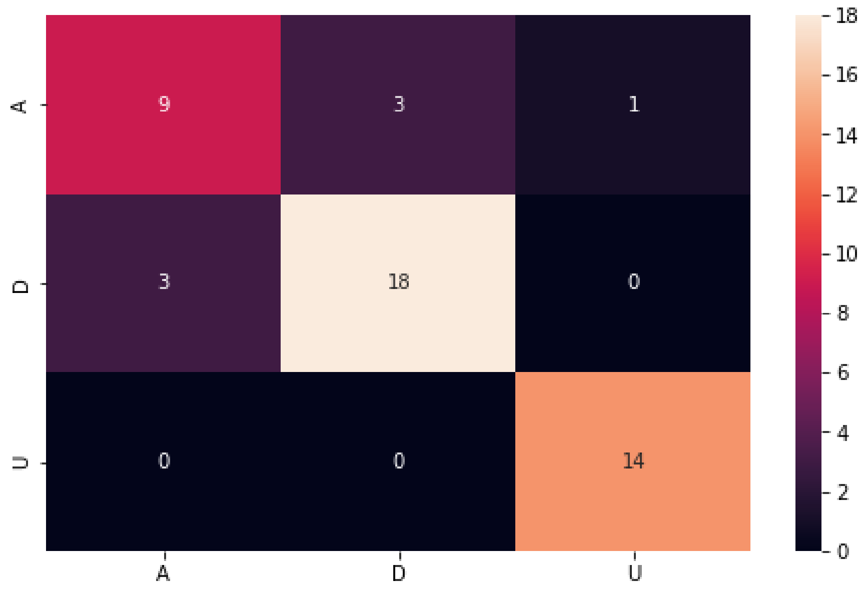


We can see from the diagonal in the confusion matrix that the Logistic Regression have predicted 8 data points as Affected, 18 data points as Dead and 10 data points as Unaffected. From the first row we can see that the Logistic Regression predicted 0 Affected data points as Dead and 5 Affected data points as Unaffected. From the second row we can see that the Logistic Regression predicted 2 Dead data points as Affected and 1 Dead data points as Unaffected. From the third row we can see that the Logistic Regression predicted 2 Unaffected data points as Affected and 2 Unaffected data points as Dead.

Confusion matrix for Decision Tree on Balanced dataset for PDT treatment:



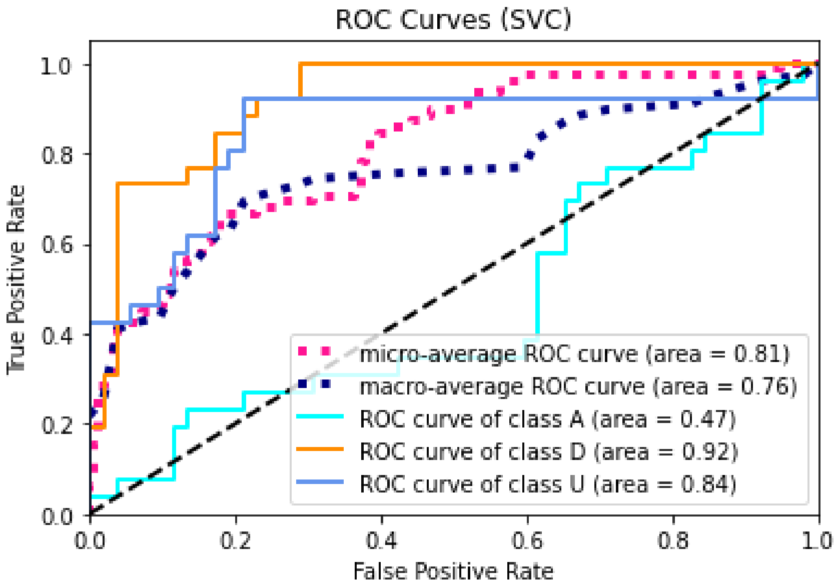
```
Decision Tree Classifier
Confusion Matrix
              precision    recall  f1-score   support

           A       0.75      0.69      0.72        13
           D       0.86      0.86      0.86        21
           U       0.93      1.00      0.97        14

    accuracy                           0.85        48
   macro avg       0.85      0.85      0.85        48
weighted avg       0.85      0.85      0.85        48
```



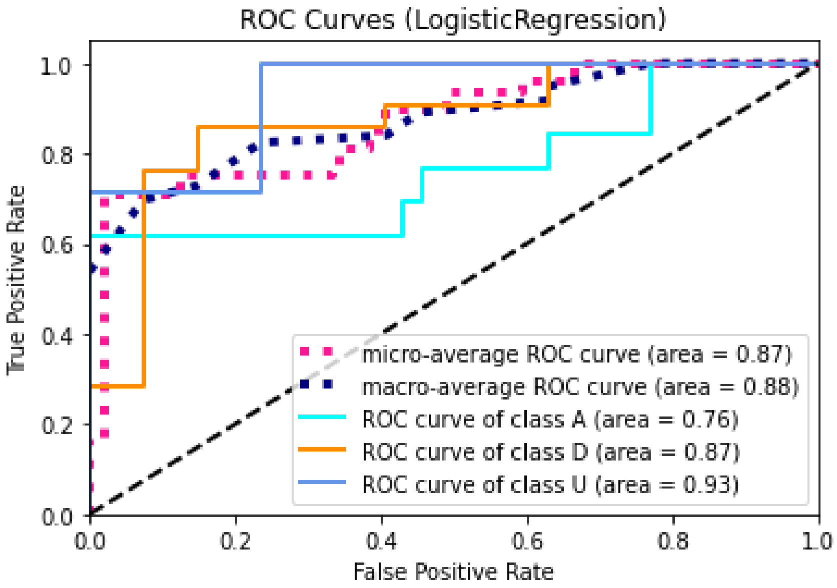
We can see from the diagonal in the confusion matrix that the Decision Tree have predicted 9 data points as Affected, 18 data points as Dead and 14 data points as Unaffected. From the first row we can see that the Decision Tree predicted 3 Affected data points as Dead and 1 Affected data points as Unaffected. From the second row we can see that the Decision Tree predicted 3 Dead data points as Affected and 0 Dead data points as Unaffected. From the third row we can see that the Decision Tree predicted 0 Unaffected data points as Affected and 0 Unaffected data points as Dead.

Now let us see the ROC Curves:

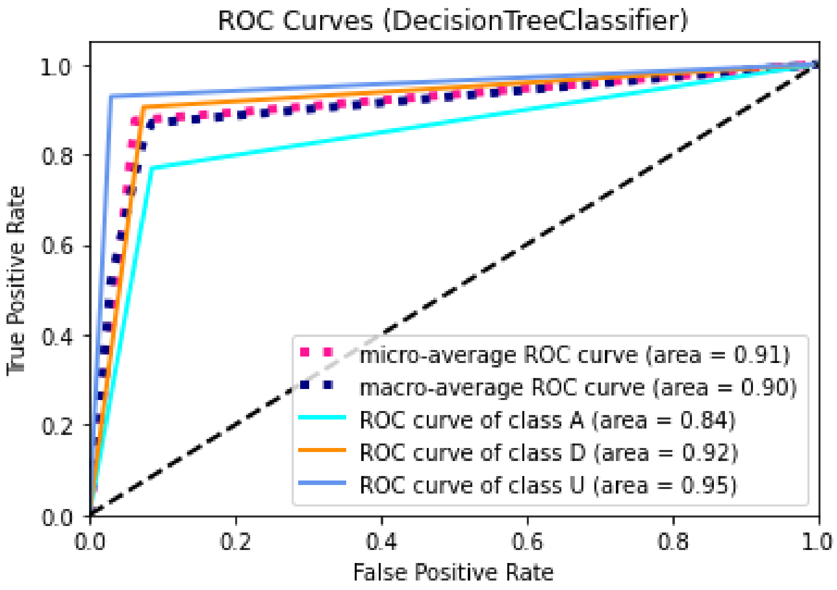
ROC curves for Linear SVM on Balanced dataset for PDT treatment:



ROC curves for Logistic Regression on Balanced dataset for PDT treatment:



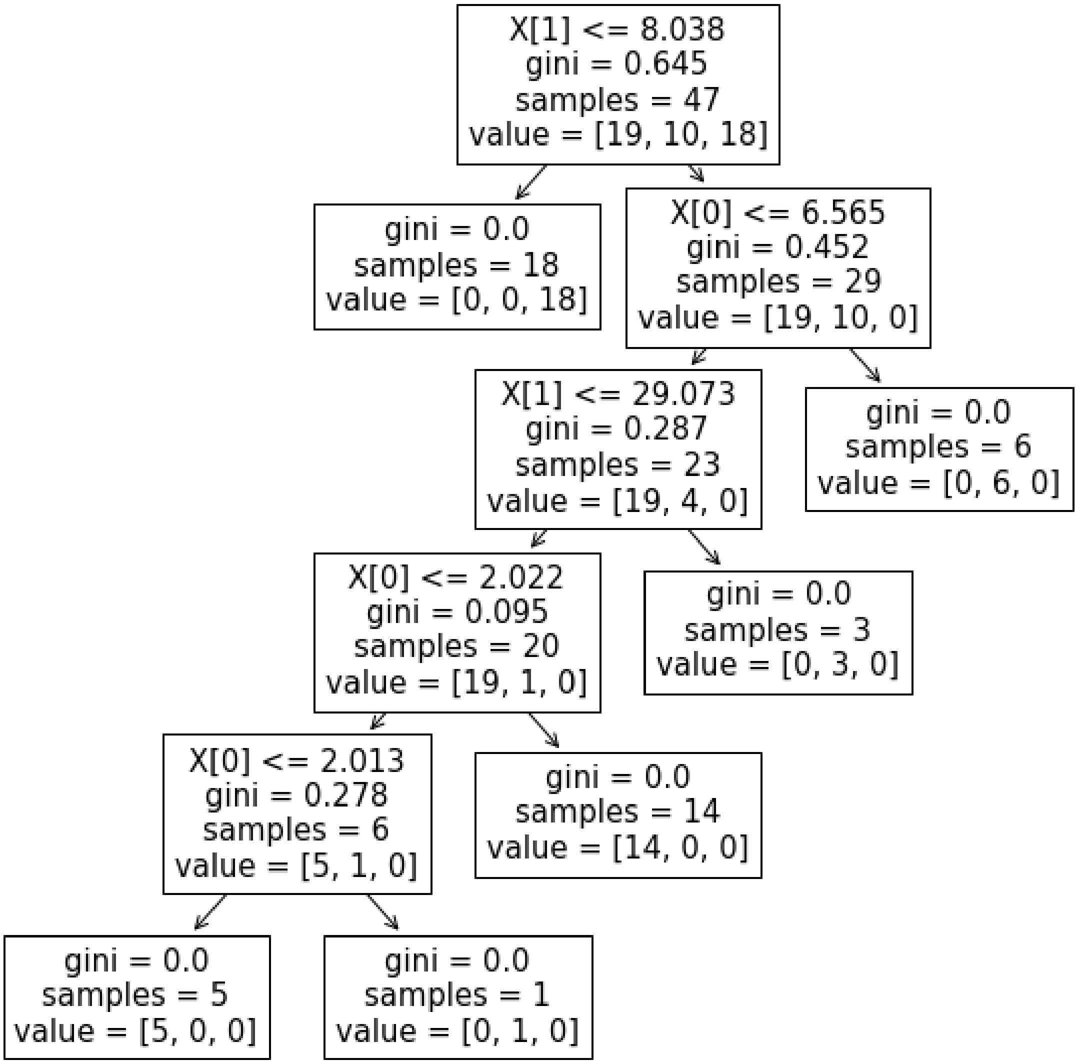
ROC curves for Decision Tree on Balanced dataset for PDT treatment:



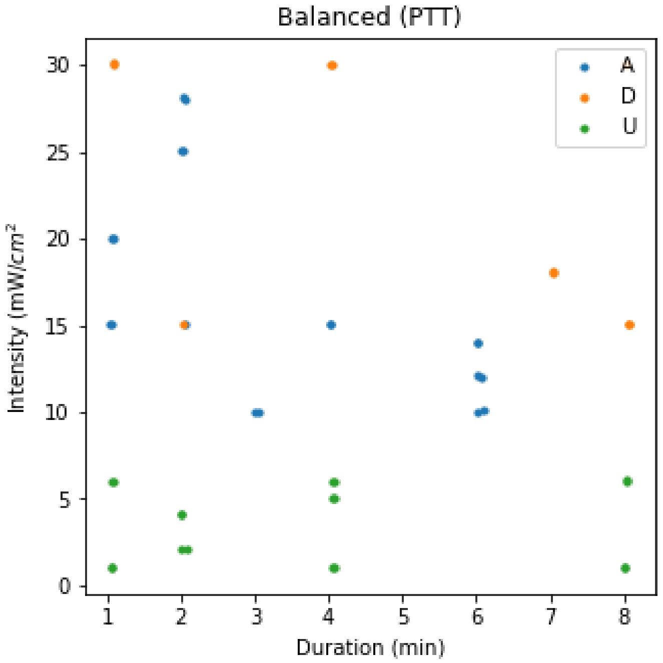
Here again we immediately see that the class A is the most difficult one to predict in general, especially for the linear models and this can be understood from the scatter plot above remembering that we are doing here a “One Vs All” approach for each class.

Looking at all our results, the decision tree seems to be the best fit for our problem for the moment. This is particularly interesting because this model is very interpretable, and we can in fact derive rules from it to classify our tumors. We show below one tree we obtained from one training during the loo procedure for PTT Treatment.

We can see in the decision tree we got for PTT treatment, figure below, if the intensity of the laser is lower than 8.038 mW/cm^2 then we should probably assign the class U. This sounds reasonable with the scatterplot presented above.


X[1] <= 8.038
gini = 0.645
samples = 47
value = [19, 10, 18]
gini = 0.0
samples = 18
value = [0, 0, 18]
X[0] <= 6.565
gini = 0.452
samples = 29
value = [19, 10, 0]
X[1] <= 29.073
gini = 0.287
samples = 23
value = [19, 4, 0]
gini = 0.0
samples = 6
value = [0, 6, 0]
X[0] <= 2.022
gini = 0.095
samples = 20
value = [19, 1, 0]
gini = 0.0
samples = 3
value = [0, 3, 0]
X[0] <= 2.013
gini = 0.278
samples = 6
value = [5, 1, 0]
gini = 0.0
samples = 14
value = [14, 0, 0]
gini = 0.0
samples = 5
value = [5, 0, 0]
gini = 0.0
samples = 1
value = [0, 1, 0]

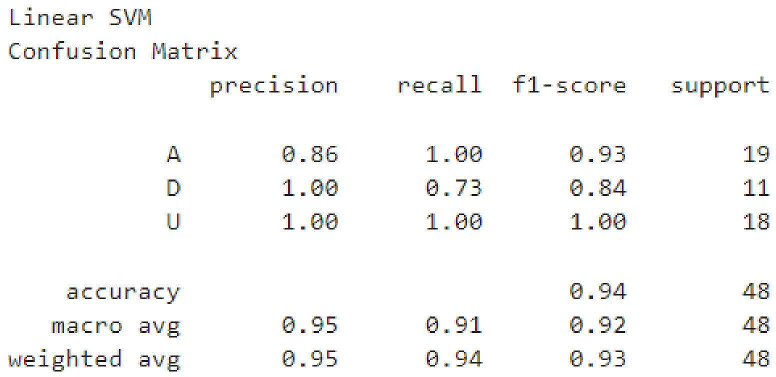
3.2 Balanced Dataset for PTT Treatment



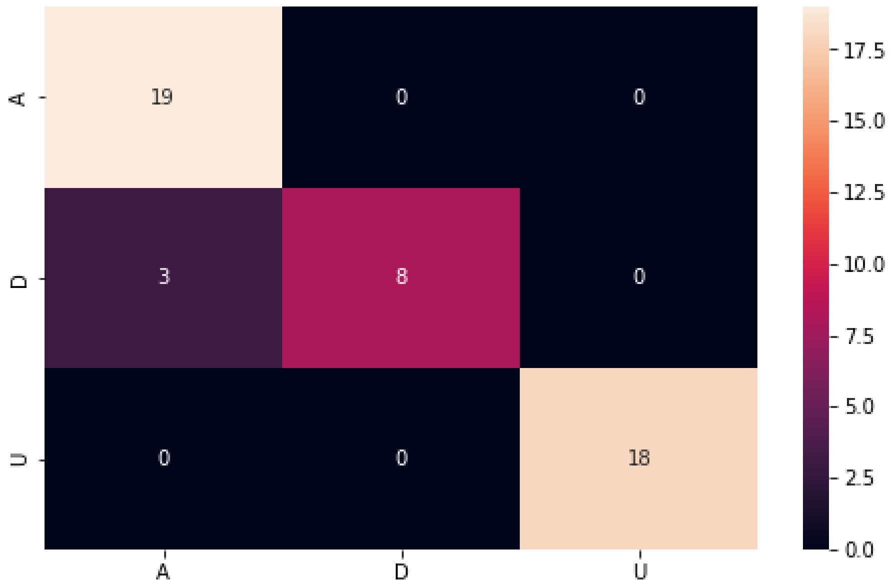
From this scatter plot we can see that the classes can be more or less linearly separable like the PDT scatter. We will use the same classifiers LinearSVM, logistic regression and decision tree. We can also guess that the class U will be easily guessed by a linear model, and the class D will not be easy since some of the points of class D is in the middle of the class A. We expect a good result in class U, bad results in class D and better results in class A.

In the balanced PTT dataset we have 48 data points from them 19 data points have the Affected label, 11 data points have the Dead label and 18 data points have the Unaffected label. In the unbalanced dataset we have 11 data points have the Affected label, 8 data points have the Dead label and 29 data points have the Unaffected label. In the SMOTE dataset we have 87 data points with 29 data points in every class.

Confusion matrix for Linear SVM on Balanced dataset for PTT treatment:

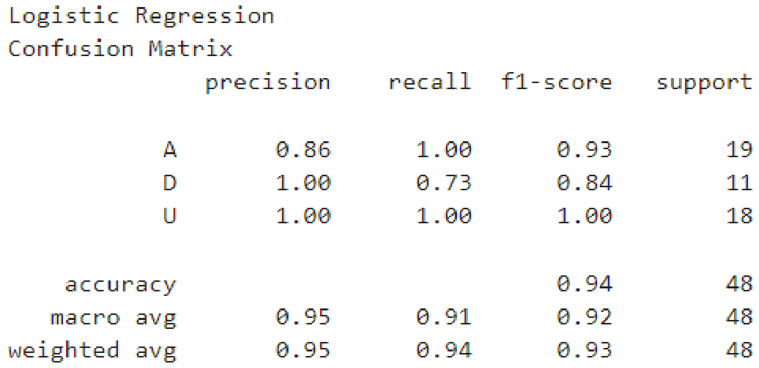


```
Linear SVM
Confusion Matrix
              precision    recall  f1-score   support

           A       0.86      1.00      0.93        19
           D       1.00      0.73      0.84        11
           U       1.00      1.00      1.00        18

    accuracy                           0.94        48
   macro avg       0.95      0.91      0.92        48
weighted avg       0.95      0.94      0.93        48
```

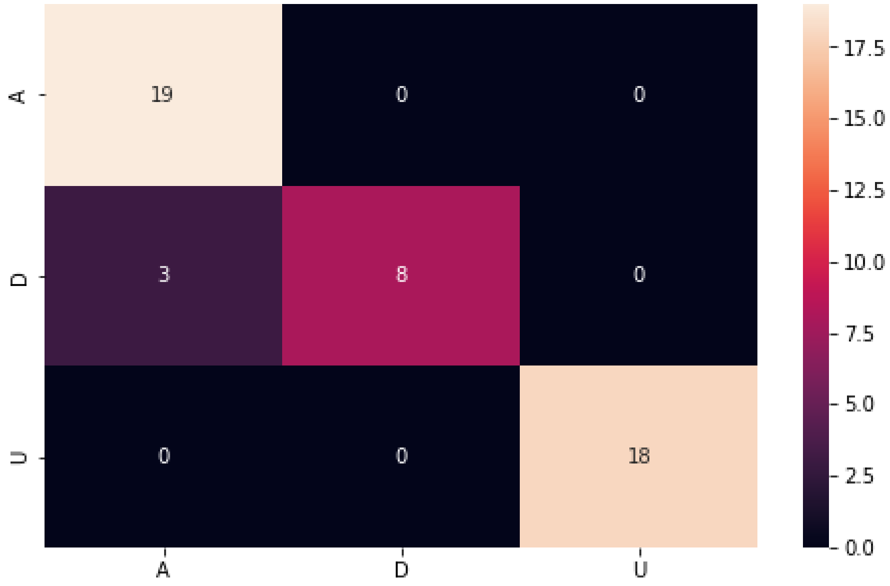


We can see from the diagonal in the confusion matrix that the Linear SVM have predicted 19 data points as Affected, 8 data points as Dead and 18 data points as Unaffected. From the first row we can see that the Linear SVM predicted 0 Affected data points as Dead and 0 Affected data points as Unaffected. From the second row we can see that the Linear SVM predicted 3 Dead data points as Affected and 0 Dead data points as Unaffected. From the third row we can see that the Linear SVM predicted 0 Unaffected data points as Affected and 0 Unaffected data points as Dead.

Confusion matrix for Logistic Regression on Balanced dataset for PTT treatment:

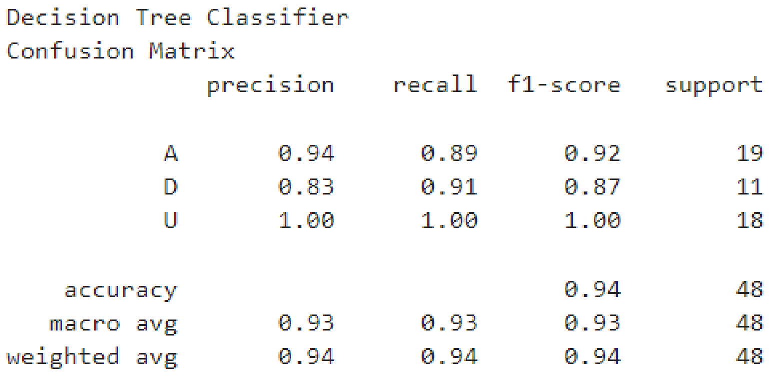


```
Logistic Regression
Confusion Matrix
              precision    recall  f1-score   support

           A       0.86      1.00      0.93        19
           D       1.00      0.73      0.84        11
           U       1.00      1.00      1.00        18

    accuracy                           0.94        48
   macro avg       0.95      0.91      0.92        48
weighted avg       0.95      0.94      0.93        48
```

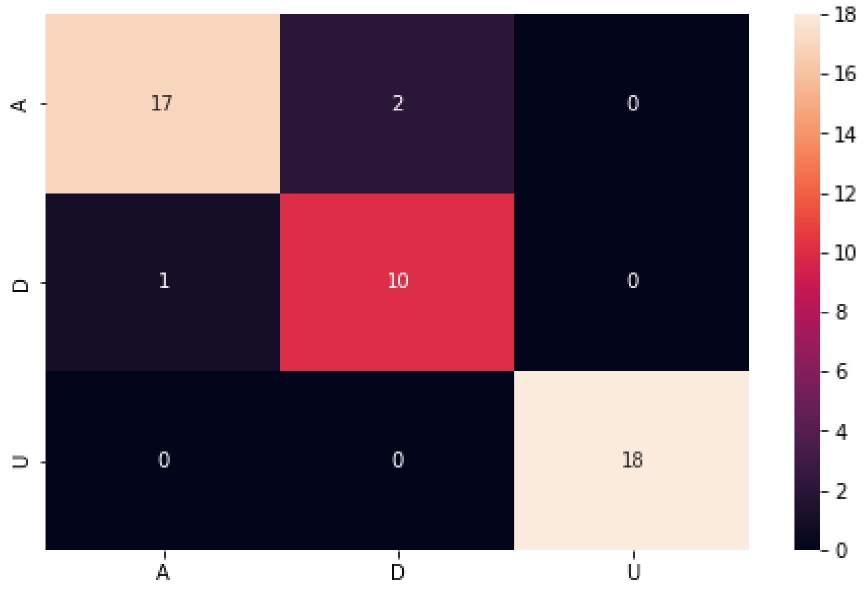


We can see from the diagonal in the confusion matrix that the Logistic Regression have predicted 19 data points as Affected, 8 data points as Dead and 18 data points as Unaffected. From the first row we can see that the Logistic Regression predicted 0 Affected data points as Dead and 0 Affected data points as Unaffected. From the second row we can see that the Logistic Regression predicted 3 Dead data points as Affected and 0 Dead data points as Unaffected. From the third row we can see that the Logistic Regression predicted 0 Unaffected data points as Affected and 0 Unaffected data points as Dead.

Confusion matrix for Decision Tree on Balanced dataset for PTT treatment:



```
Decision Tree Classifier
Confusion Matrix
              precision    recall  f1-score   support

           A       0.94      0.89      0.92        19
           D       0.83      0.91      0.87        11
           U       1.00      1.00      1.00        18

    accuracy                           0.94        48
   macro avg       0.93      0.93      0.93        48
weighted avg       0.94      0.94      0.94        48
```

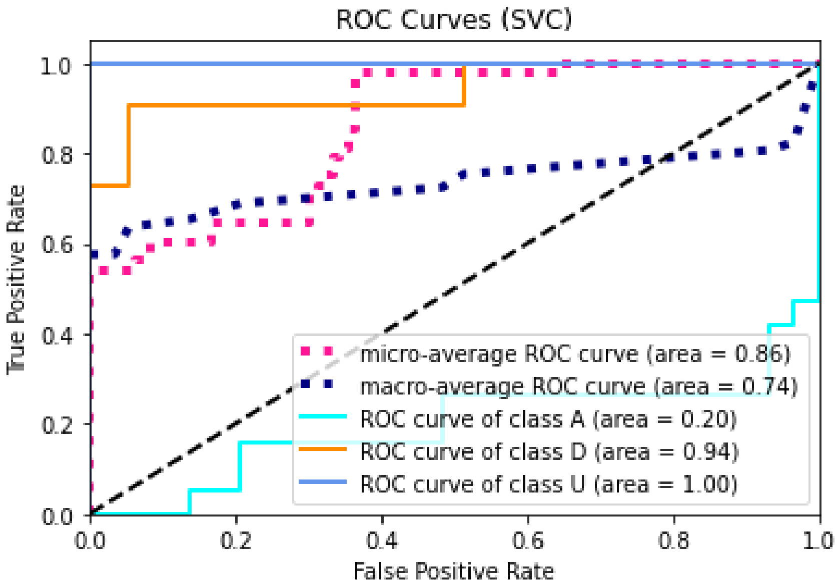


We can see from the diagonal in the confusion matrix that the Decision Tree have predicted 17 data points as Affected, 10 data points as Dead and 18 data points as Unaffected. From the first row we can see that the Decision Tree predicted 2 Affected data points as Dead and 0 Affected data points as Unaffected. From the second row we can see that the Decision Tree predicted 1 Dead data points as Affected and 0 Dead data points as Unaffected. From the third row we can see that the Decision Tree predicted 0 Unaffected data points as Affected and 0 Unaffected data points as Dead.

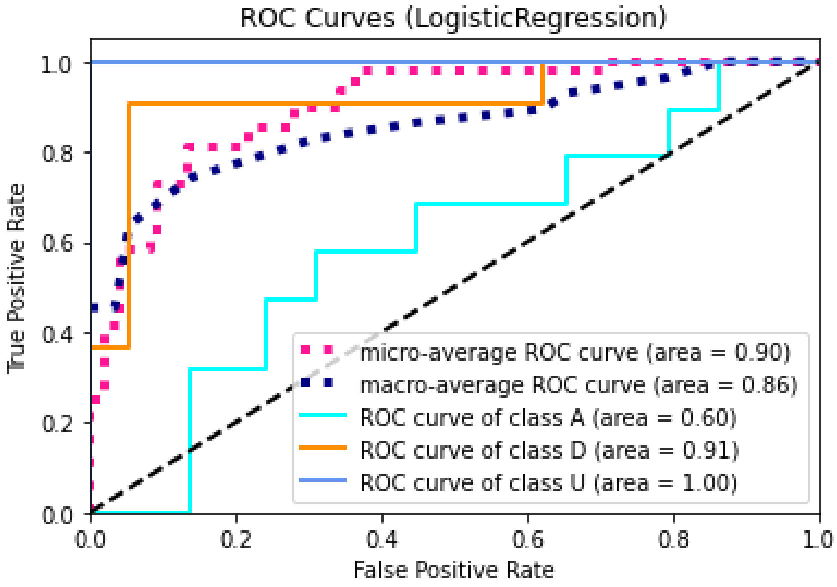
As we can see, all the models are performing reasonably well. The most difficult class to predict is the class D as expected.

Now let us see the ROC Curves:

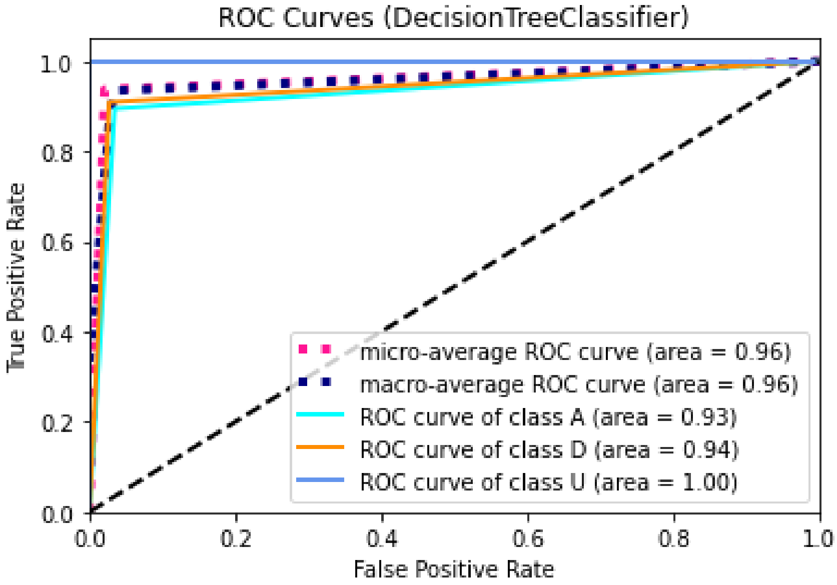
ROC curves for Linear SVM on Balanced dataset for PTT treatment:



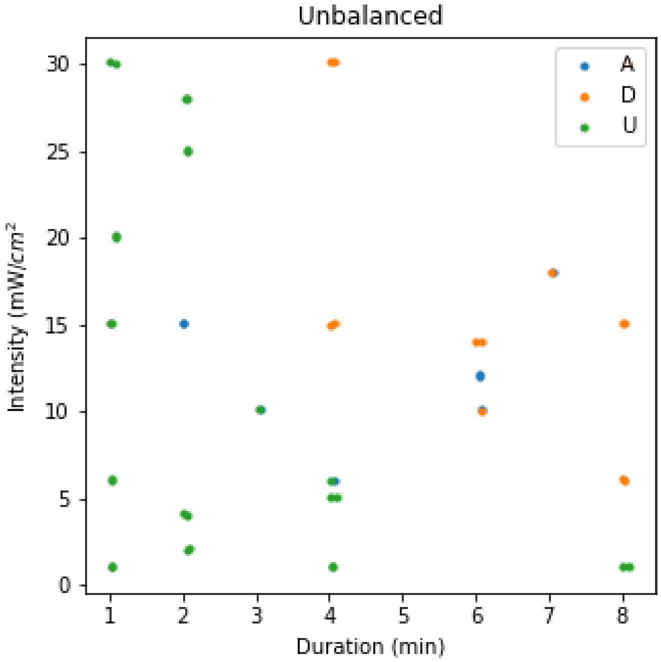
ROC curves for Logistic Regression on Balanced dataset for PTT treatment:



ROC curves for Decision Tree on Balanced dataset for PTT treatment:

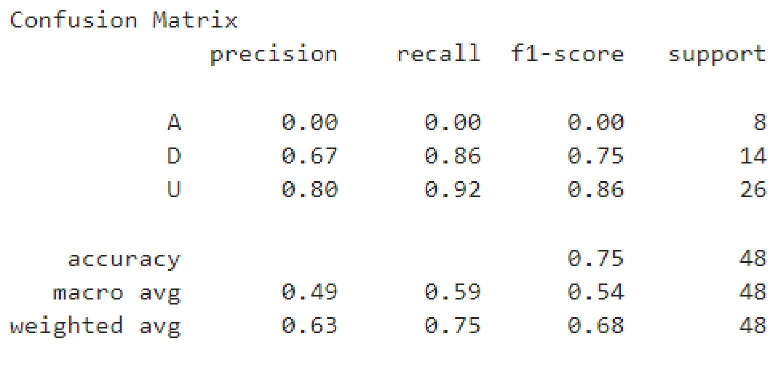


3.3 Unbalanced Dataset for PDT Treatment

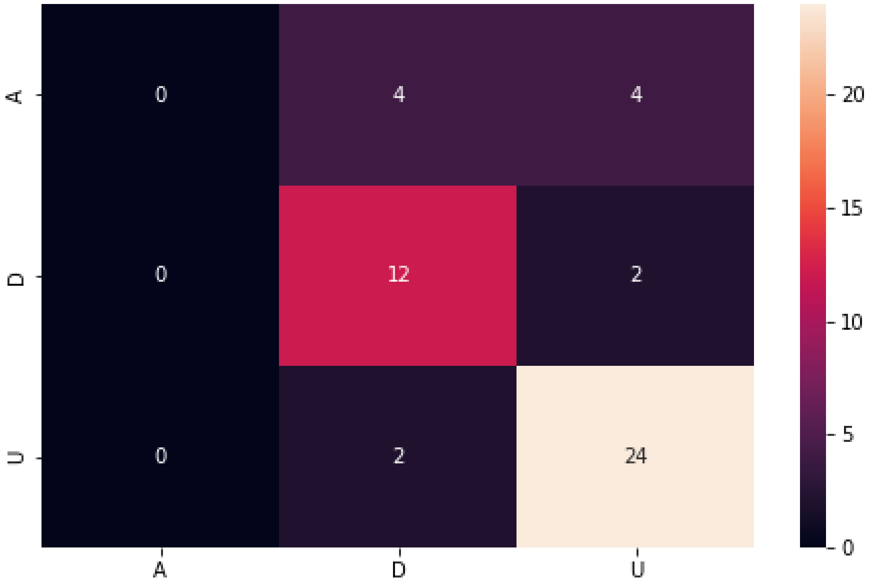


From this scatter plot we can see that the D and U classes can be more or less linearly separable. We can also guess that the class A will not be easy to separate from both classes. We expect a good result in class U, bad results in class A and better results in class D.

Confusion matrix for Linear SVM on Unbalanced dataset for PDT treatment:

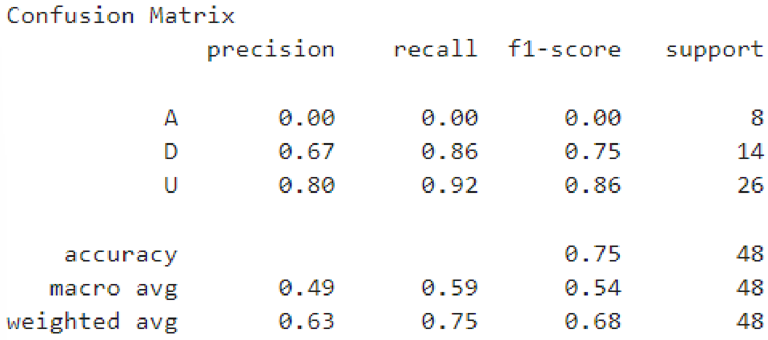


```
Confusion Matrix
              precision    recall  f1-score   support

           A       0.00      0.00      0.00         8
           D       0.67      0.86      0.75        14
           U       0.80      0.92      0.86        26

    accuracy                           0.75        48
   macro avg       0.49      0.59      0.54        48
weighted avg       0.63      0.75      0.68        48
```

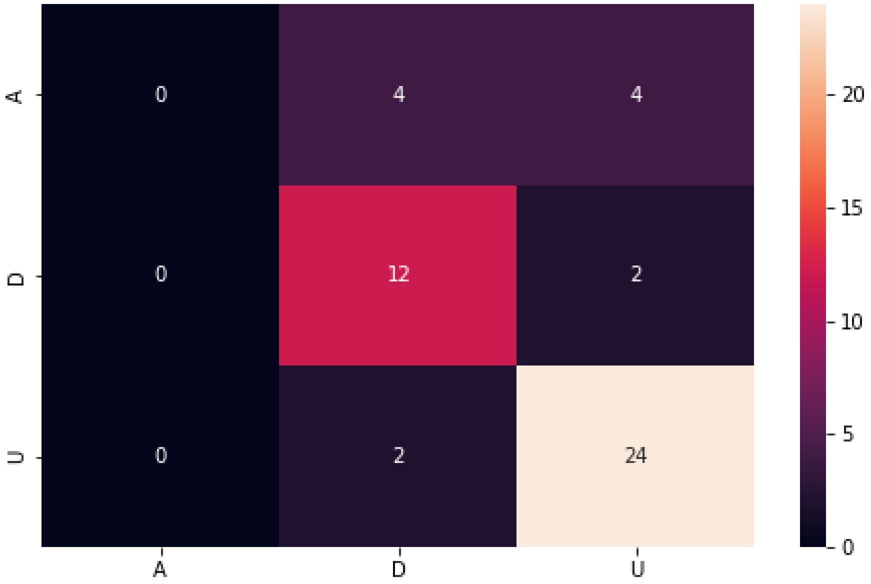


We can see from the diagonal in the confusion matrix that the Linear SVM have predicted 0 data points as Affected, 12 data points as Dead and 24 data points as Unaffected. From the first row we can see that the Linear SVM predicted 4 Affected data points as Dead and 4 Affected data points as Unaffected. From the second row we can see that the Linear SVM predicted 0 Dead data points as Affected and 2 Dead data points as Unaffected. From the third row we can see that the Linear SVM predicted 0 Unaffected data points as Affected and 2 Unaffected data points as Dead.

Confusion matrix for Logistic Regression on Unbalanced dataset for PDT treatment:

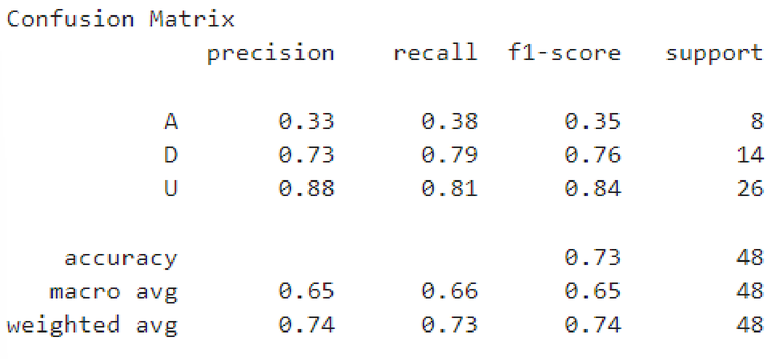


```
Confusion Matrix
              precision    recall  f1-score   support

           A       0.00      0.00      0.00         8
           D       0.67      0.86      0.75        14
           U       0.80      0.92      0.86        26

    accuracy                           0.75        48
   macro avg       0.49      0.59      0.54        48
weighted avg       0.63      0.75      0.68        48
```

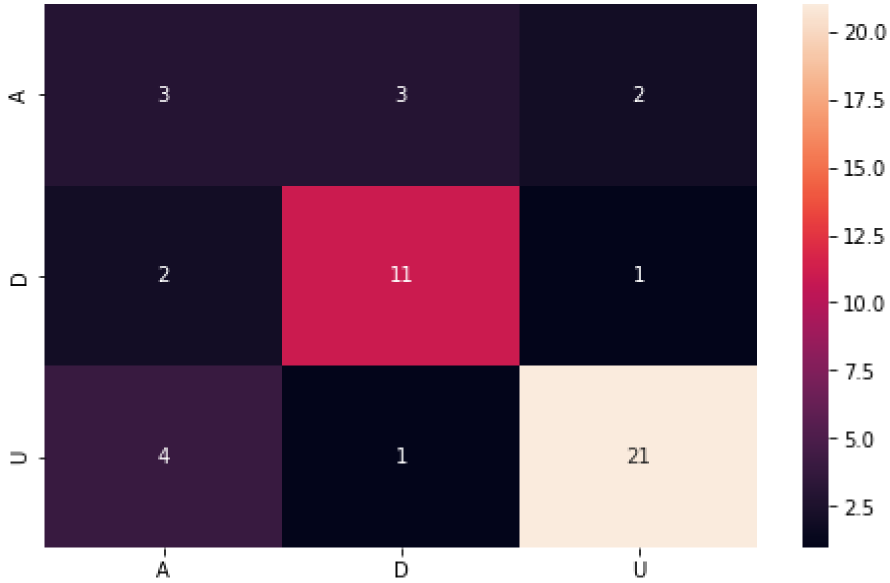


We can see from the diagonal in the confusion matrix that the Logistic Regression have predicted 0 data points as Affected, 12 data points as Dead and 24 data points as Unaffected. From the first row we can see that the Logistic Regression predicted 0 Affected data points as Dead and 5 Affected data points as Unaffected. From the second row we can see that the Logistic Regression predicted 0 Dead data points as Affected and 2 Dead data points as Unaffected. From the third row we can see that the Logistic Regression predicted 0 Unaffected data points as Affected and 2 Unaffected data points as Dead.

Confusion matrix for Decision Tree on Unbalanced dataset for PDT treatment:



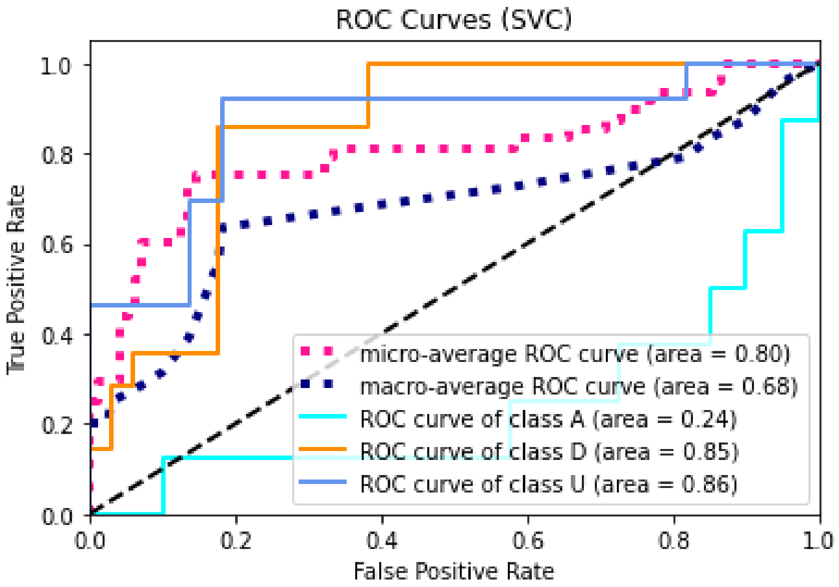
```
Confusion Matrix
              precision    recall  f1-score   support

           A       0.33      0.38      0.35         8
           D       0.73      0.79      0.76        14
           U       0.88      0.81      0.84        26

    accuracy                           0.73        48
   macro avg       0.65      0.66      0.65        48
weighted avg       0.74      0.73      0.74        48
```



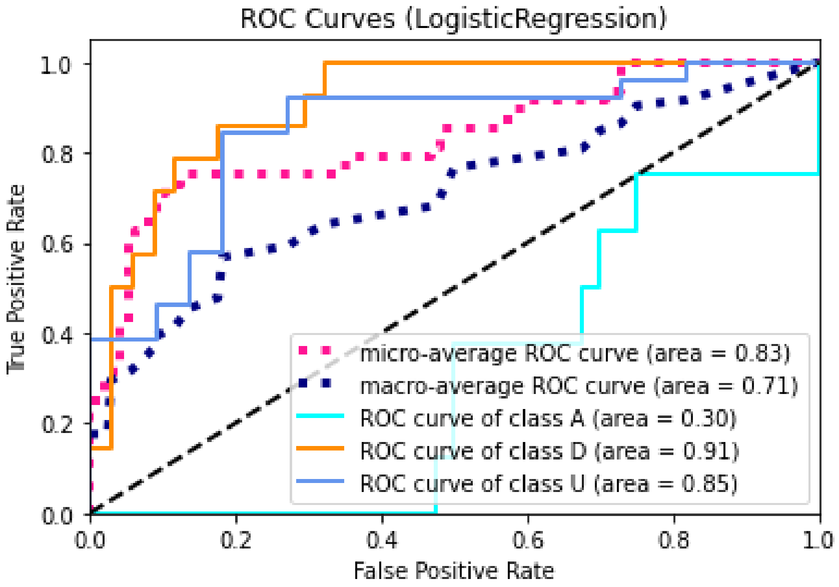
We can see from the diagonal in the confusion matrix that the Decision Tree have predicted 3 data points as Affected, 11 data points as Dead and 21 data points as Unaffected. From the first row we can see that the Decision Tree predicted 3 Affected data points as Dead and 2 Affected data points as Unaffected. From the second row we can see that the Decision Tree predicted 2 Dead data points as Affected and 1 Dead data points as Unaffected. From the third row we can see that the Decision Tree predicted 4 Unaffected data points as Affected and 1 Unaffected data points as Dead.

Now let us see the ROC Curves:

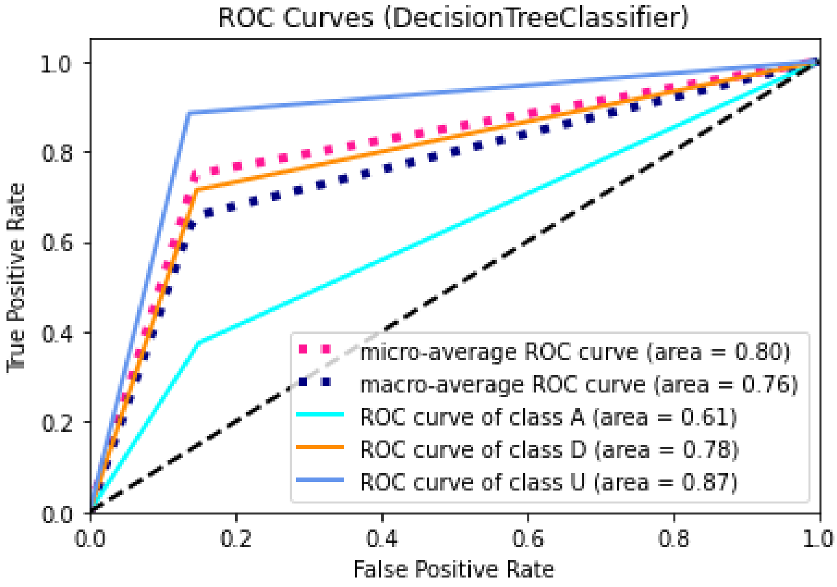
ROC curves for Linear SVM on Unbalanced dataset for PDT treatment:



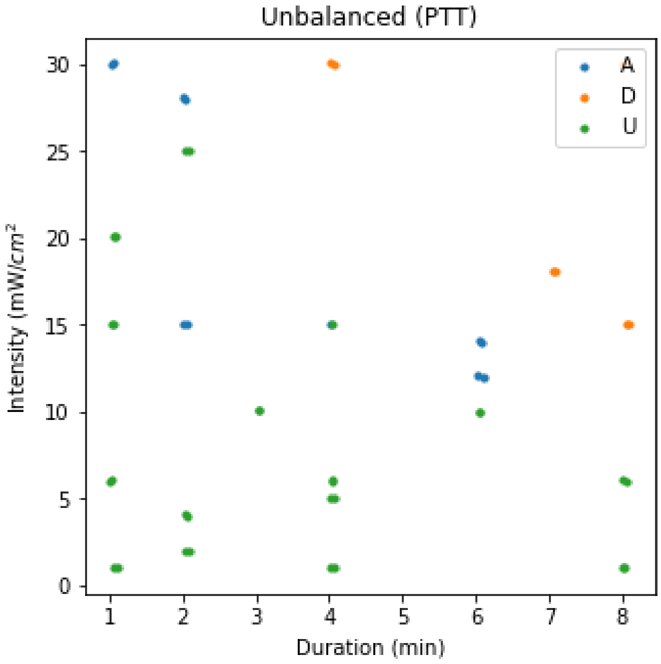
ROC curves for Logistic Regression on Unbalanced dataset for PDT treatment:



ROC curves for Decision Tree on Unbalanced dataset for PDT treatment:

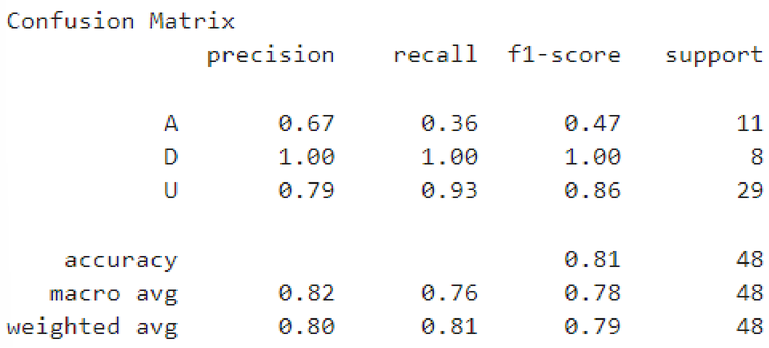


3.4 Unbalanced Dataset for PTT Treatment

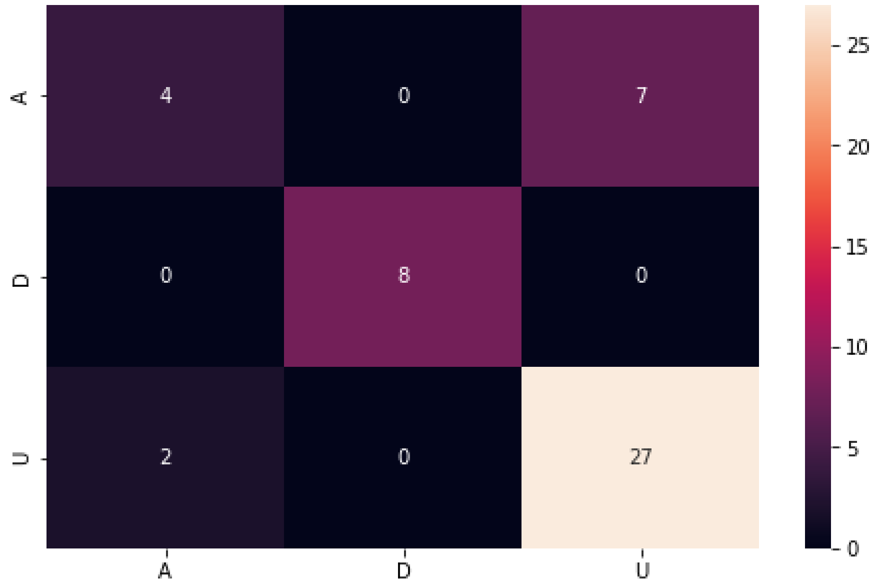


From this scatter plot we can see that the classes can be more or less linearly separable. We can also guess that the class D will not be easy to separate from both classes. We expect a good result in class D, bad results in class A and better results in class U.

Confusion matrix for Linear SVM on Unbalanced dataset for PTT treatment:

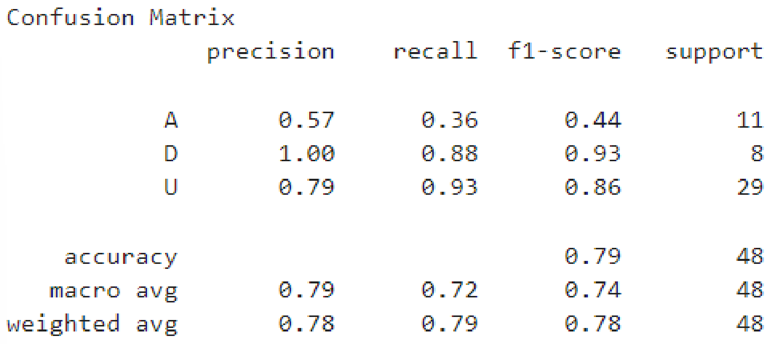


```
Confusion Matrix
              precision    recall  f1-score   support

           A       0.67      0.36      0.47        11
           D       1.00      1.00      1.00         8
           U       0.79      0.93      0.86        29

    accuracy                           0.81        48
   macro avg       0.82      0.76      0.78        48
weighted avg       0.80      0.81      0.79        48
```

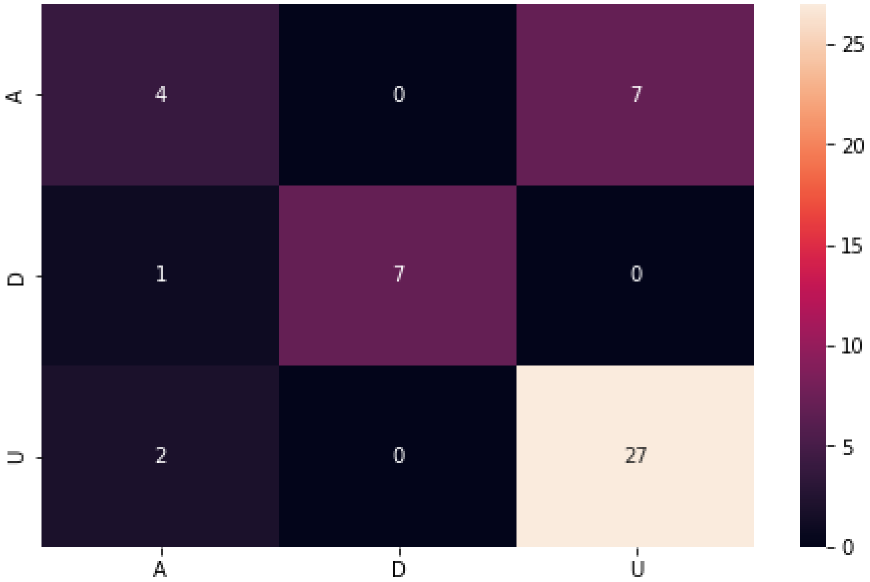


We can see from the diagonal in the confusion matrix that the Linear SVM have predicted 4 data points as Affected, 8 data points as Dead and 27 data points as Unaffected. From the first row we can see that the Linear SVM predicted 0 Affected data points as Dead and 7 Affected data points as Unaffected. From the second row we can see that the Linear SVM predicted 0 Dead data points as Affected and 0 Dead data points as Unaffected. From the third row we can see that the Linear SVM predicted 2 Unaffected data points as Affected and 0 Unaffected data points as Dead.

Confusion matrix for Logistic Regression on Unbalanced dataset for PTT treatment:

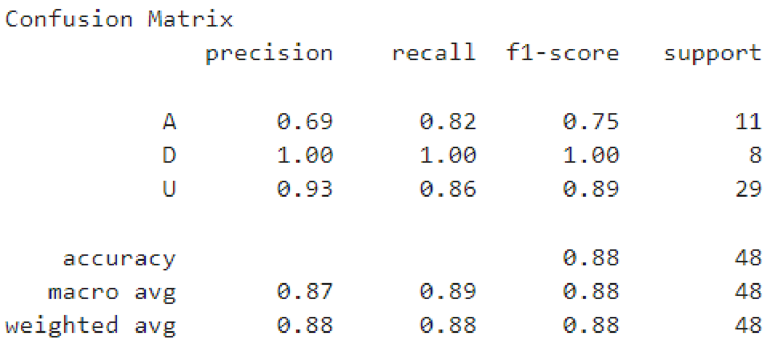


```
Confusion Matrix
              precision    recall  f1-score   support

           A       0.57      0.36      0.44        11
           D       1.00      0.88      0.93         8
           U       0.79      0.93      0.86        29

    accuracy                           0.79        48
   macro avg       0.79      0.72      0.74        48
weighted avg       0.78      0.79      0.78        48
```

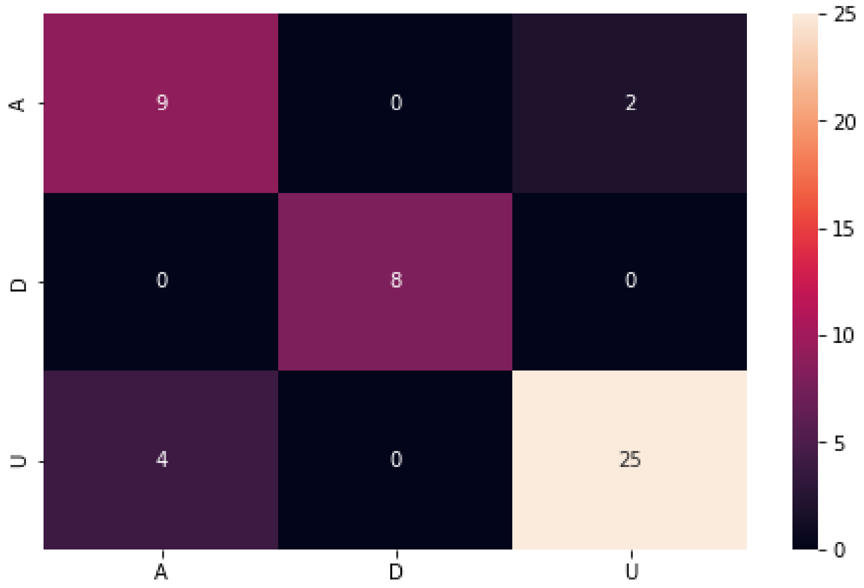


We can see from the diagonal in the confusion matrix that the Logistic Regression have predicted 4 data points as Affected, 7 data points as Dead and 27 data points as Unaffected. From the first row we can see that the Logistic Regression predicted 0 Affected data points as Dead and 7 Affected data points as Unaffected. From the second row we can see that the Logistic Regression predicted 1 Dead data points as Affected and 0 Dead data points as Unaffected. From the third row we can see that the Logistic Regression predicted 2 Unaffected data points as Affected and 0 Unaffected data points as Dead.

Confusion matrix for Decision Tree on Unbalanced dataset for PTT treatment:



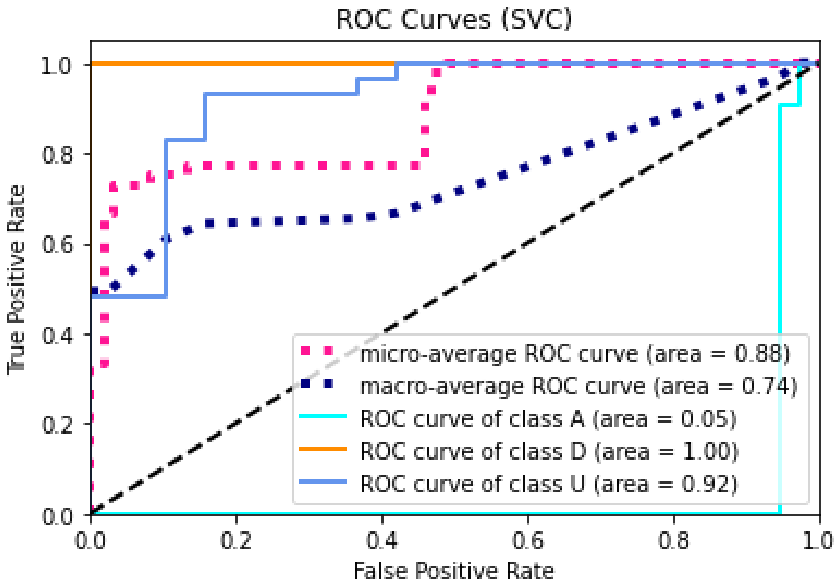
```
Confusion Matrix
              precision    recall  f1-score   support

           A       0.69      0.82      0.75        11
           D       1.00      1.00      1.00         8
           U       0.93      0.86      0.89        29

    accuracy                           0.88        48
   macro avg       0.87      0.89      0.88        48
weighted avg       0.88      0.88      0.88        48
```



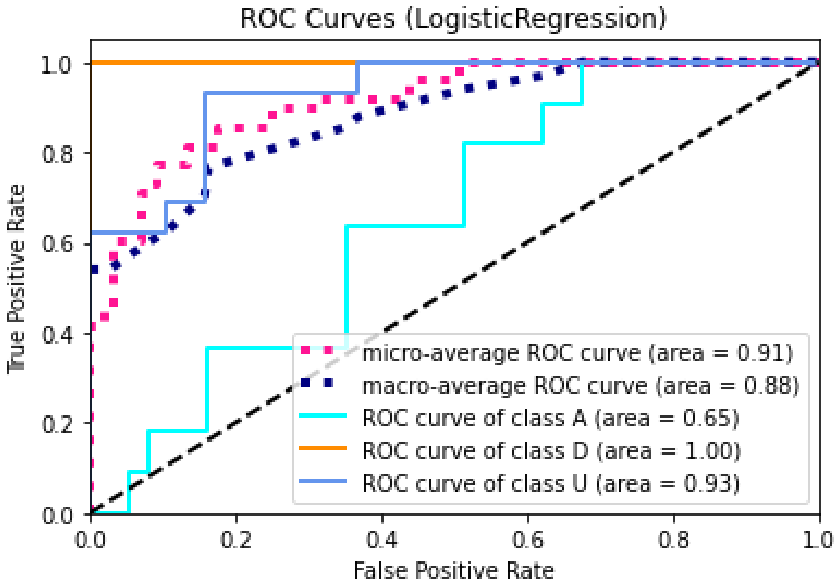
We can see from the diagonal in the confusion matrix that the Decision Tree have predicted 9 data points as Affected, 8 data points as Dead and 25 data points as Unaffected. From the first row we can see that the Decision Tree predicted 0 Affected data points as Dead and 2 Affected data points as Unaffected. From the second row we can see that the Decision Tree predicted 0 Dead data points as Affected and 0 Dead data points as Unaffected. From the third row we can see that the Decision Tree predicted 4 Unaffected data points as Affected and 0 Unaffected data points as Dead.

Now let us see the ROC Curves:

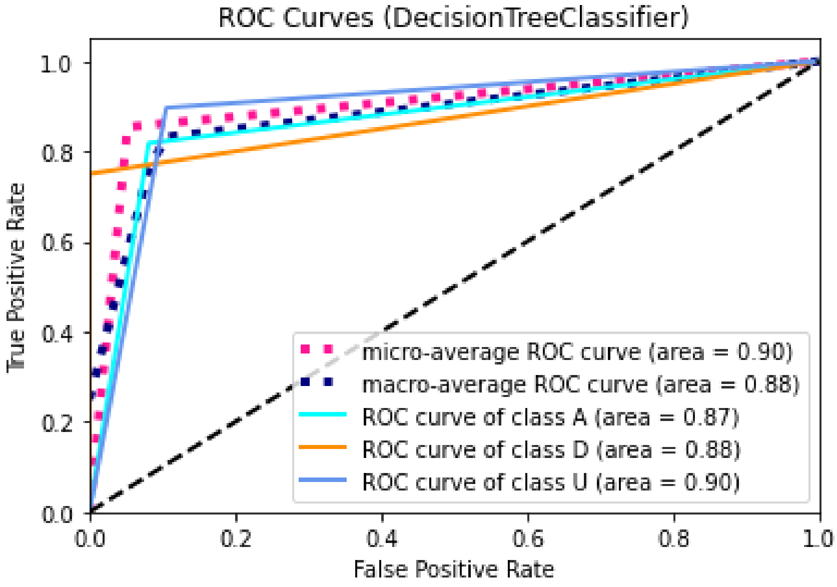
ROC curves for Linear SVM on Unbalanced dataset for PTT treatment:



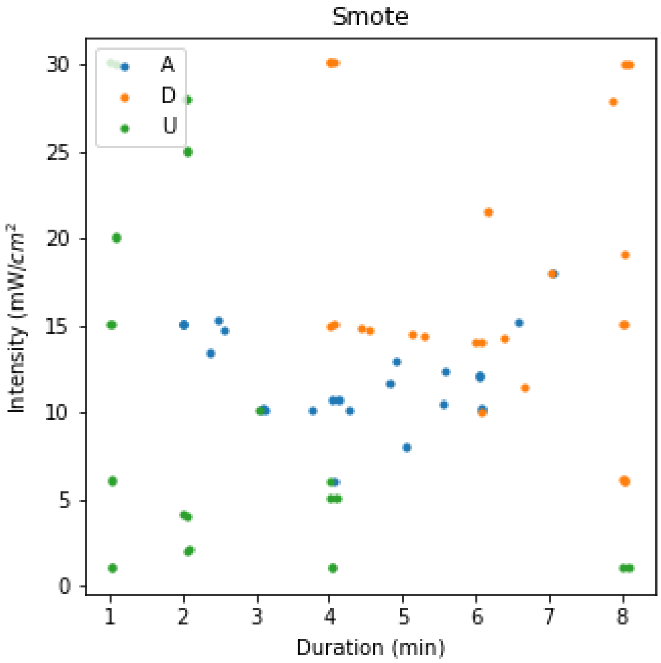
ROC curves for Logistic Regression on Unbalanced dataset for PTT treatment:



ROC curves for Decision Tree on Unbalanced dataset for PTT treatment:

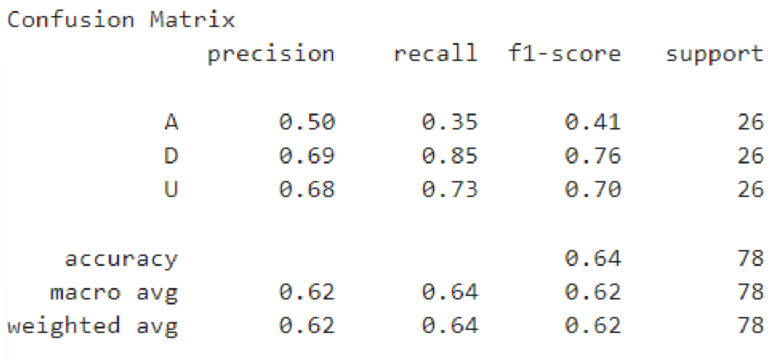


3.5 SMOTE Dataset for PDT Treatment

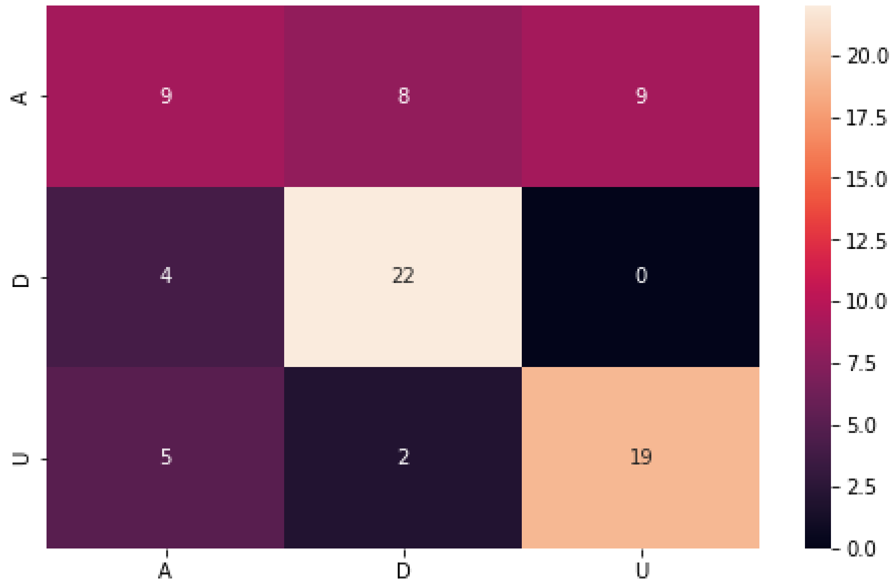


From this scatter plot we can see that the classes are less linearly separable, and we assume that the decision tree will have better results than Linear SVM and logistic regression.

Confusion matrix for Linear SVM on SMOTE dataset for PDT treatment:

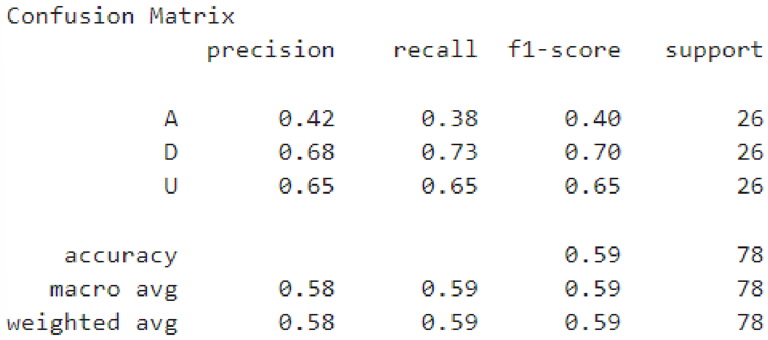


```
Confusion Matrix
              precision    recall  f1-score   support

           A       0.50      0.35      0.41        26
           D       0.69      0.85      0.76        26
           U       0.68      0.73      0.70        26

    accuracy                           0.64        78
   macro avg       0.62      0.64      0.62        78
weighted avg       0.62      0.64      0.62        78
```

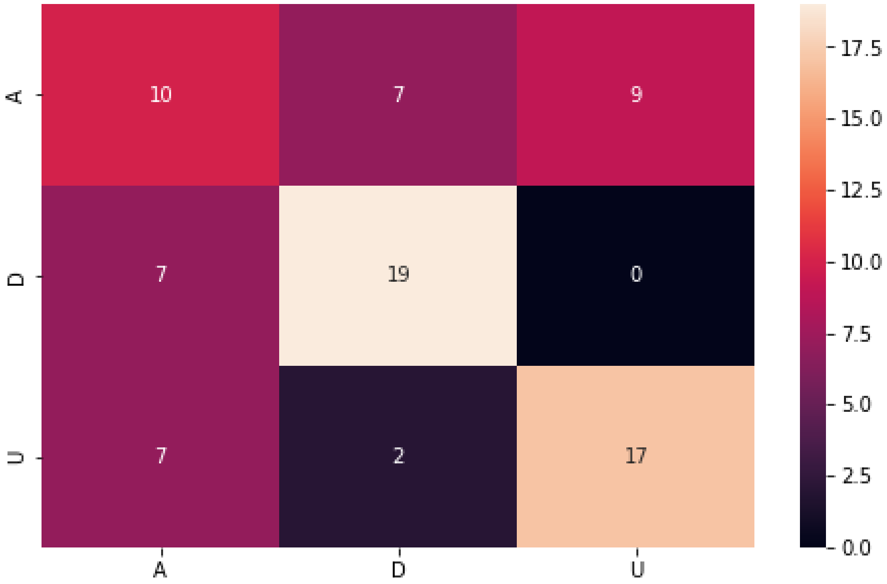


We can see from the diagonal in the confusion matrix that the Linear SVM have predicted 9 data points as Affected, 22 data points as Dead and 19 data points as Unaffected. From the first row we can see that the Linear SVM predicted 8 Affected data points as Dead and 9 Affected data points as Unaffected. From the second row we can see that the Linear SVM predicted 4 Dead data points as Affected and 0 Dead data points as Unaffected. From the third row we can see that the Linear SVM predicted 5 Unaffected data points as Affected and 2 Unaffected data points as Dead.

Confusion matrix for Logistic Regression on SMOTE dataset for PDT treatment:

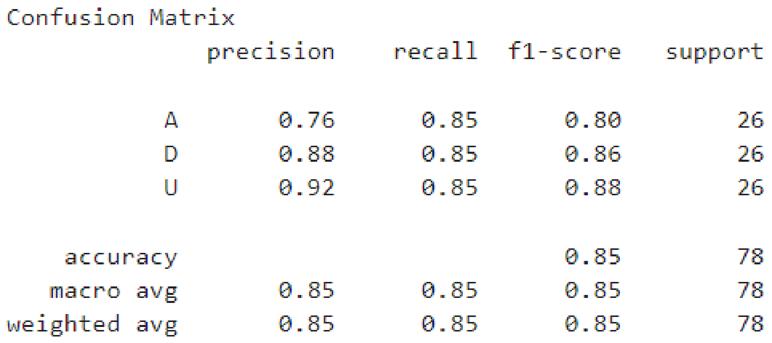


```
Confusion Matrix
              precision    recall  f1-score   support

           A       0.42      0.38      0.40        26
           D       0.68      0.73      0.70        26
           U       0.65      0.65      0.65        26

    accuracy                           0.59        78
   macro avg       0.58      0.59      0.59        78
weighted avg       0.58      0.59      0.59        78
```

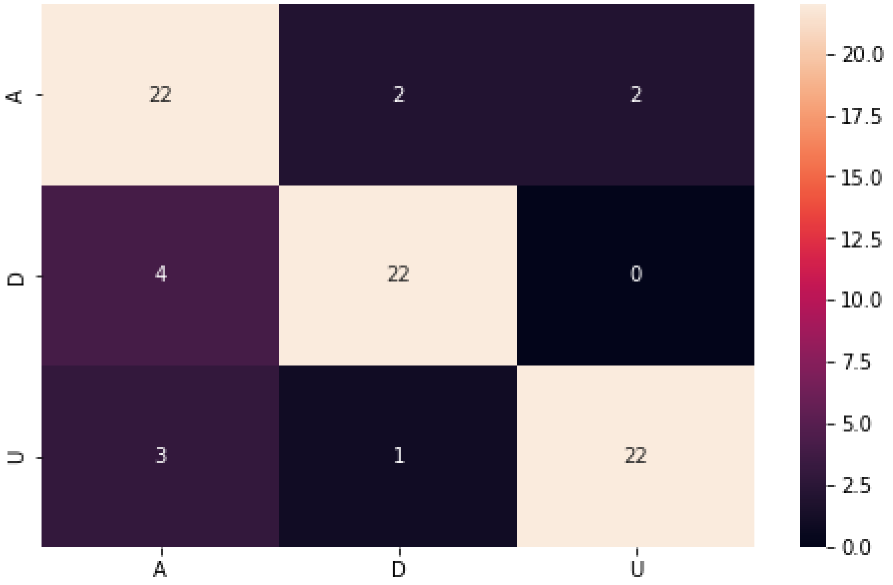


We can see from the diagonal in the confusion matrix that the Logistic Regression have predicted 10 data points as Affected, 19 data points as Dead and 17 data points as Unaffected. From the first row we can see that the Logistic Regression predicted 7 Affected data points as Dead and 9 Affected data points as Unaffected. From the second row we can see that the Logistic Regression predicted 7 Dead data points as Affected and 0 Dead data points as Unaffected. From the third row we can see that the Logistic Regression predicted 7 Unaffected data points as Affected and 2 Unaffected data points as Dead.

Confusion matrix for Decision Tree on SMOTE dataset for PDT treatment:



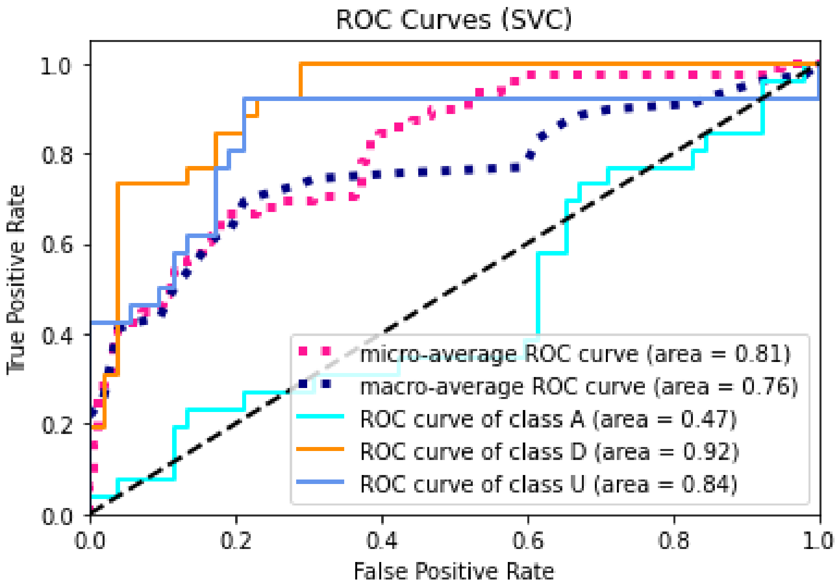
```
Confusion Matrix
              precision    recall  f1-score   support

           A       0.76      0.85      0.80        26
           D       0.88      0.85      0.86        26
           U       0.92      0.85      0.88        26

    accuracy                           0.85        78
   macro avg       0.85      0.85      0.85        78
weighted avg       0.85      0.85      0.85        78
```



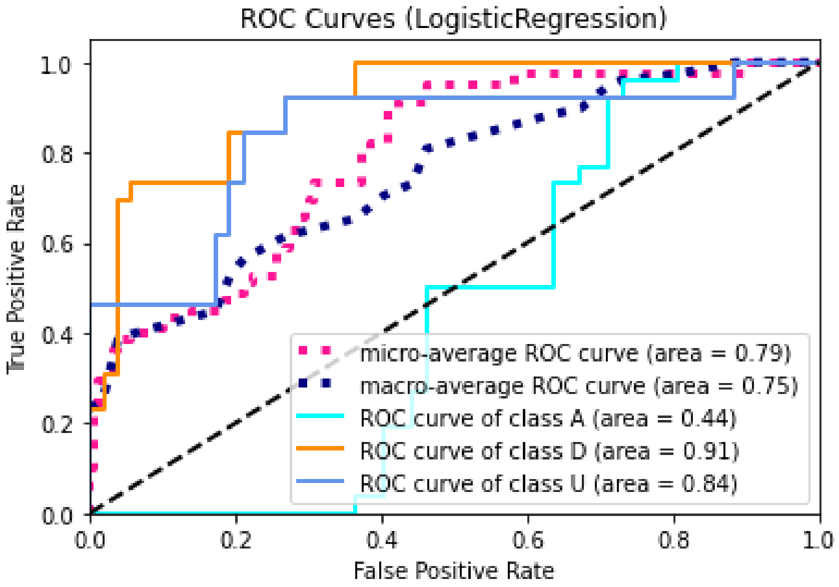
We can see from the diagonal in the confusion matrix that the Decision Tree have predicted 22 data points as Affected, 22 data points as Dead and 22 data points as Unaffected. From the first row we can see that the Decision Tree predicted 2 Affected data points as Dead and 2 Affected data points as Unaffected. From the second row we can see that the Decision Tree predicted 4 Dead data points as Affected and 0 Dead data points as Unaffected. From the third row we can see that the Decision Tree predicted 3 Unaffected data points as Affected and 1 Unaffected data points as Dead.

Now let us see the ROC Curves:

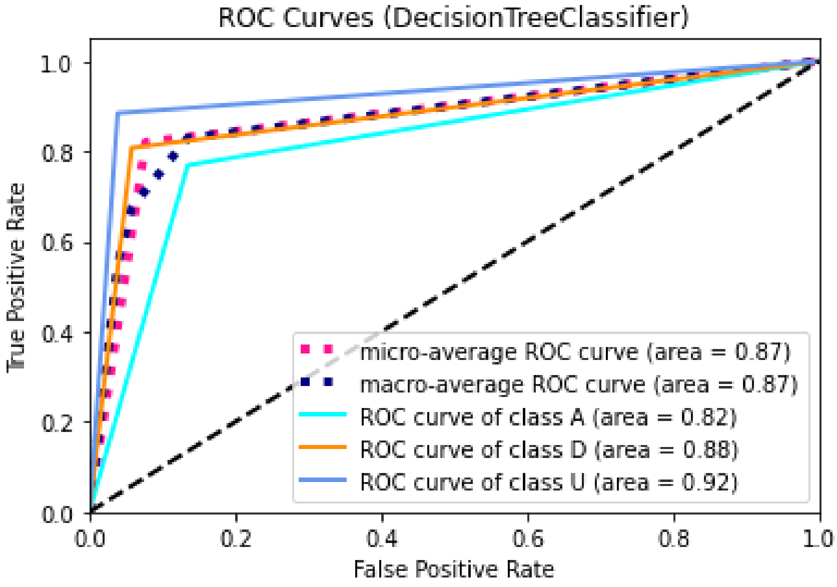
ROC curves for Linear SVM on SMOTE dataset for PDT treatment:



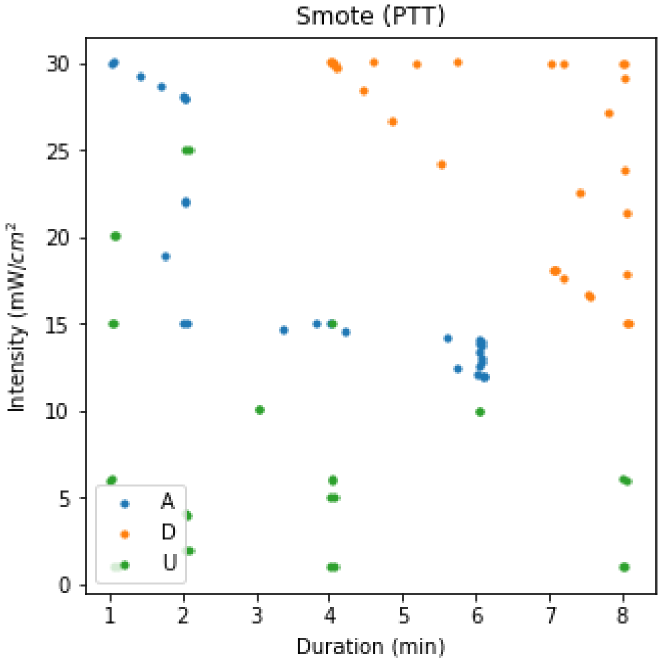
ROC curves for Logistic Regression on SMOTE dataset for PDT treatment:



ROC curves for Decision Tree on SMOTE dataset for PDT treatment:

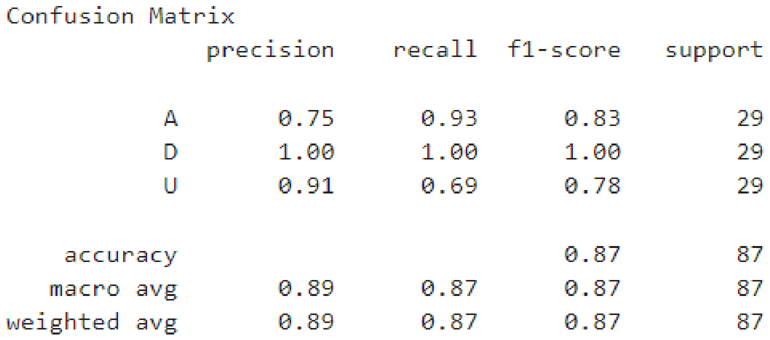


3.6 SMOTE Dataset for PTT Treatment

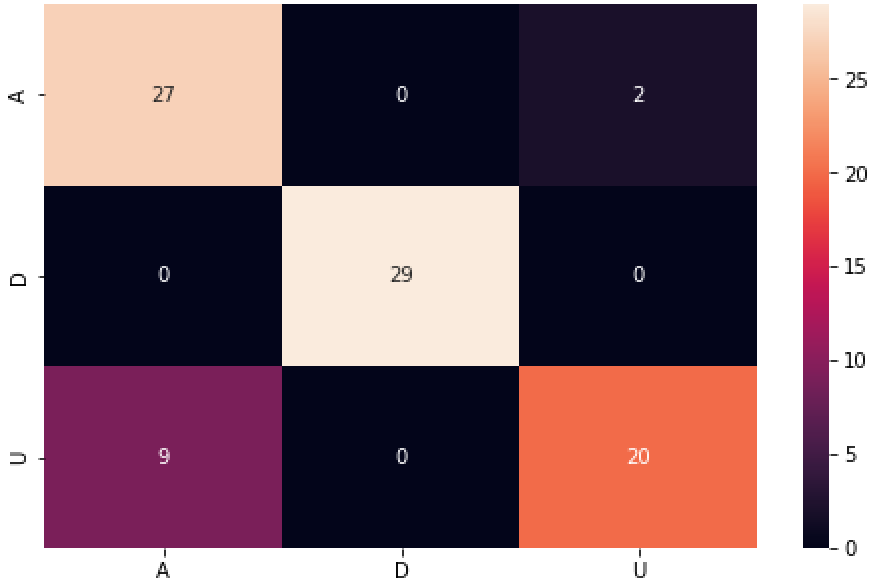


From this scatter plot we can see that the classes can be more or less linearly separable. We expect a good result in class D, worse results in classes A and D.

Confusion matrix for Linear SVM on SMOTE dataset for PTT treatment:

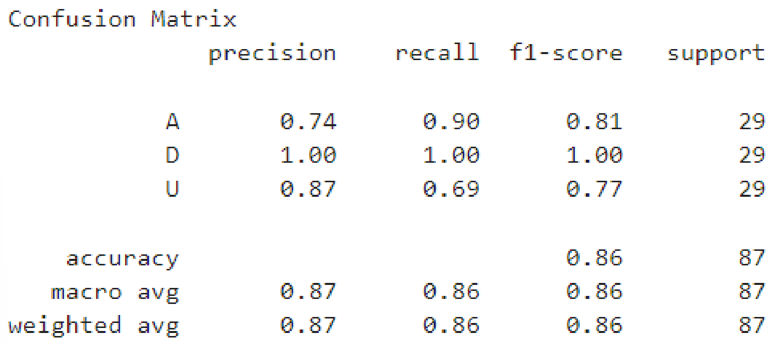


```
Confusion Matrix
              precision    recall  f1-score   support

           A       0.75      0.93      0.83        29
           D       1.00      1.00      1.00        29
           U       0.91      0.69      0.78        29

    accuracy                           0.87        87
   macro avg       0.89      0.87      0.87        87
weighted avg       0.89      0.87      0.87        87
```

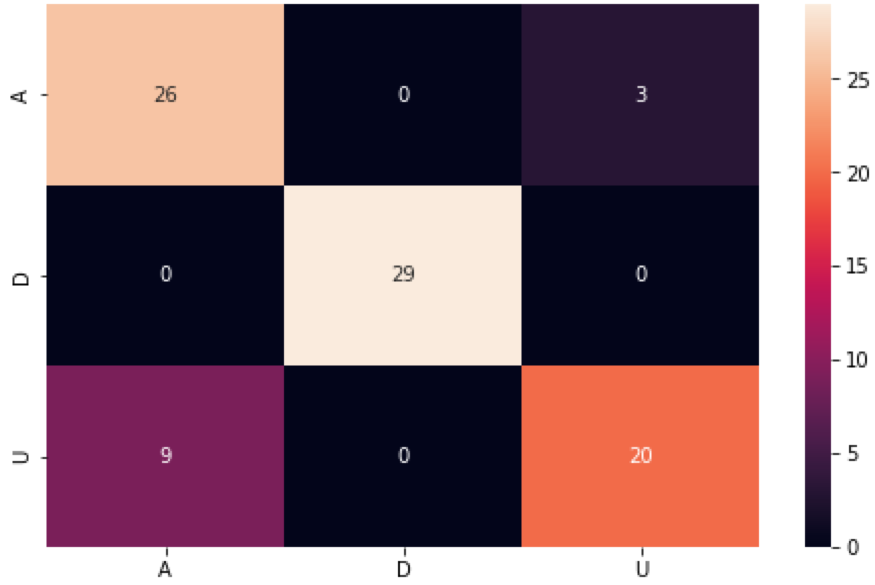


We can see from the diagonal in the confusion matrix that the Linear SVM have predicted 27 data points as Affected, 29 data points as Dead and 20 data points as Unaffected. From the first row we can see that the Linear SVM predicted 0 Affected data points as Dead and 2 Affected data points as Unaffected. From the second row we can see that the Linear SVM predicted 0 Dead data points as Affected and 0 Dead data points as Unaffected. From the third row we can see that the Linear SVM predicted 9 Unaffected data points as Affected and 0 Unaffected data points as Dead.

Confusion matrix for Logistic Regression on SMOTE dataset for PTT treatment:

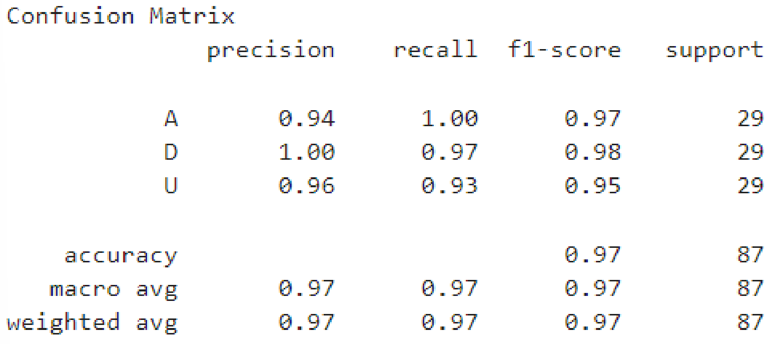


```
Confusion Matrix
              precision    recall  f1-score   support

           A       0.74      0.90      0.81        29
           D       1.00      1.00      1.00        29
           U       0.87      0.69      0.77        29

    accuracy                           0.86        87
   macro avg       0.87      0.86      0.86        87
weighted avg       0.87      0.86      0.86        87
```

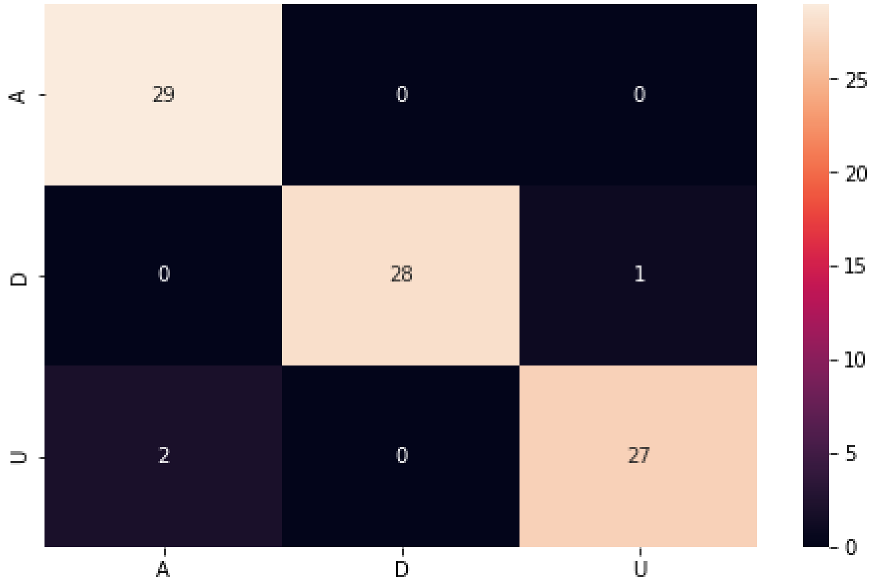


We can see from the diagonal in the confusion matrix that the Logistic Regression have predicted 26 data points as Affected, 29 data points as Dead and 20 data points as Unaffected. From the first row we can see that the Logistic Regression predicted 0 Affected data points as Dead and 3 Affected data points as Unaffected. From the second row we can see that the Logistic Regression predicted 0 Dead data points as Affected and 0 Dead data points as Unaffected. From the third row we can see that the Logistic Regression predicted 9 Unaffected data points as Affected and 0 Unaffected data points as Dead.

Confusion matrix for Decision Tree on SMOTE dataset for PTT treatment:



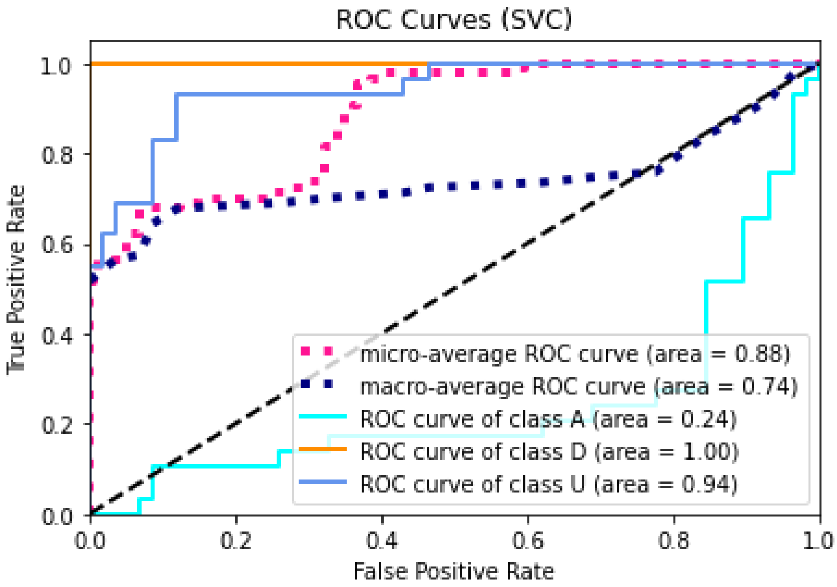
```
Confusion Matrix
              precision    recall  f1-score   support

           A       0.94      1.00      0.97        29
           D       1.00      0.97      0.98        29
           U       0.96      0.93      0.95        29

    accuracy                           0.97        87
   macro avg       0.97      0.97      0.97        87
weighted avg       0.97      0.97      0.97        87
```



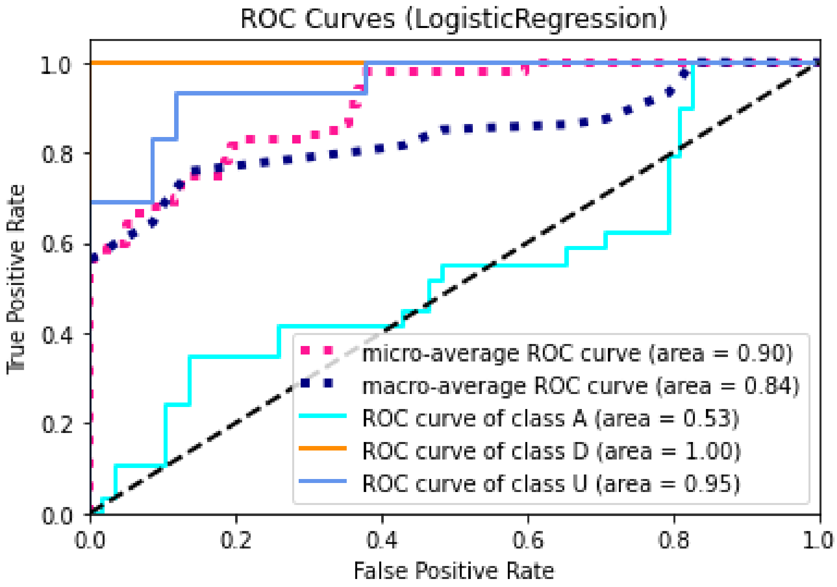
We can see from the diagonal in the confusion matrix that the Decision Tree have predicted 29 data points as Affected, 28 data points as Dead and 27 data points as Unaffected. From the first row we can see that the Decision Tree predicted 0 Affected data points as Dead and 0 Affected data points as Unaffected. From the second row we can see that the Decision Tree predicted 0 Dead data points as Affected and 1 Dead data points as Unaffected. From the third row we can see that the Decision Tree predicted 2 Unaffected data points as Affected and 0 Unaffected data points as Dead.

Now let us see the ROC Curves:

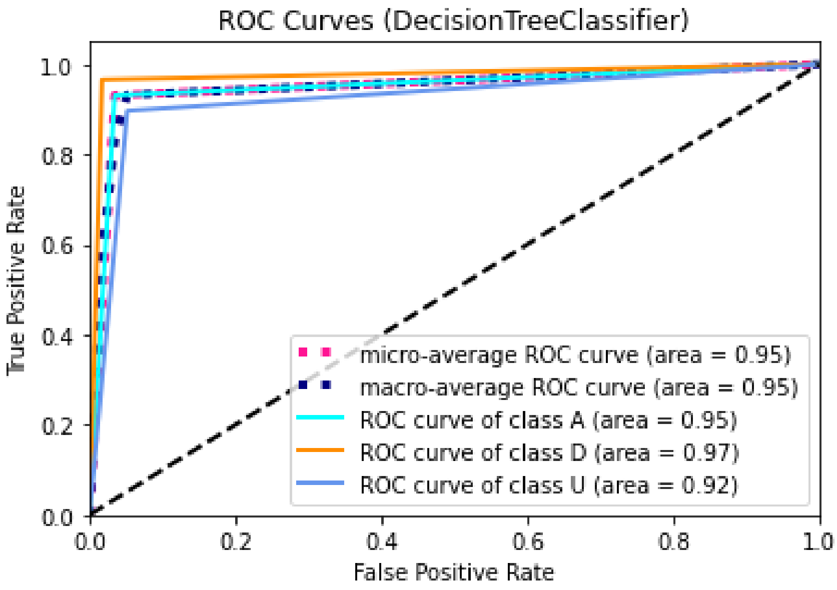
ROC curves for Linear SVM on SMOTE dataset for PTT treatment:



ROC curves for Logistic Regression on SMOTE dataset for PTT treatment:

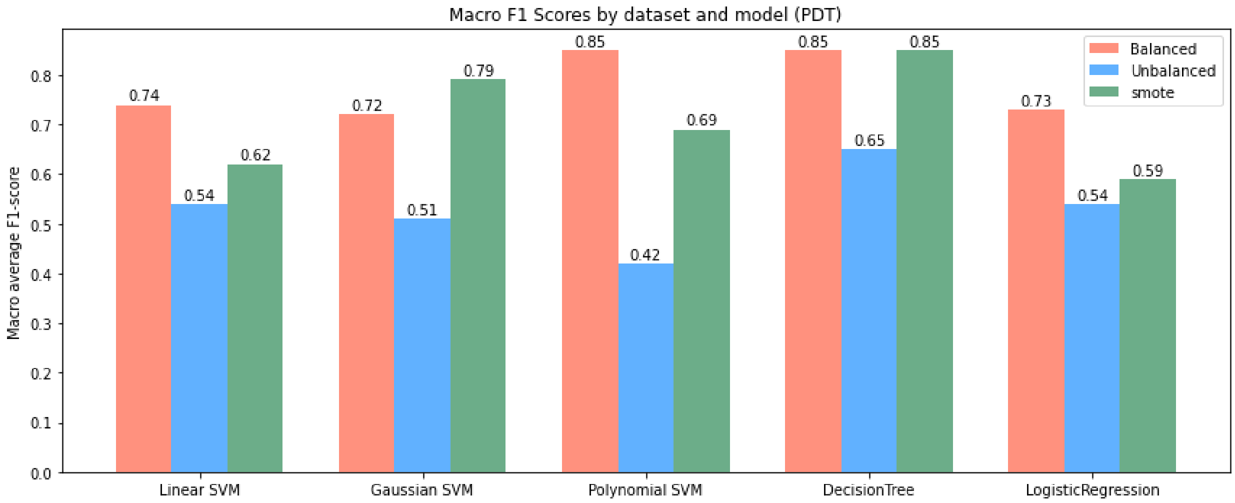


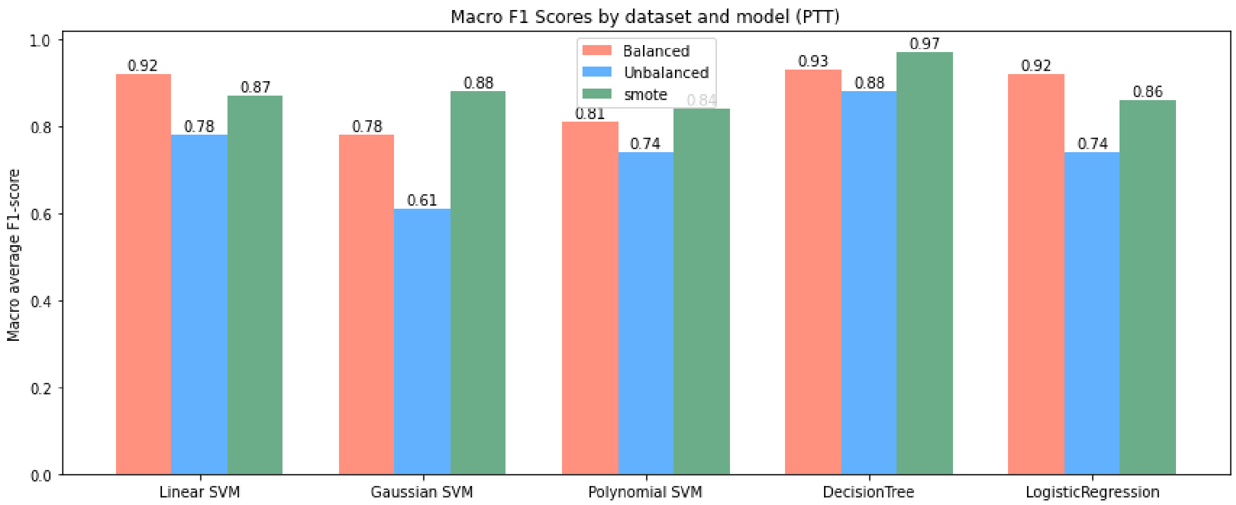
ROC curves for Decision Tree on SMOTE dataset for PTT treatment:



Comparison between the models

We presented a lot of results focusing on both treatments. We propose a simple comparison presenting the mean of the f1-scores obtained for each class in each case.





According to the way the classes are Unbalanced, the performances obtained are clearly different. This makes sense since having minority classes will foster our models to simply never predict them since they are not appearing very often.

Moreover, using the SMOTE strategy to oversample the Unbalanced dataset seems worthy since we systematically observe an improvement in terms of performance for each classifier.

The decision tree seems the best model so far for our predictions. This is interesting considering that this model is very interpretable.

Conclusions and Future Idea

In this project we tried to determine which machine learning model will predict the mortality in cancerous cell with the best results. We found that in both PDT and PTT we get the best overall results with the Decision tree model. In addition to the better accuracy results the decision tree is also good for interpretability since it gives us rules, we can follow to get our desired mortality results.

We also saw that when we used SMOTE on our dataset the performance of the models where better than those with the unbalance dataset.

As for future work we would like to test the models on other thresholds for the classes and on a bigger dataset. We can also add to the dataset different features, for example tumor cell size, and experiment combining PDT and PTT treatments on the same cells.

References

[1] "What Is Cancer?," *National Cancer Institute*, May 05, 2021. https://www.cancer.gov/about-cancer/understanding/what-is-cancer (accessed Mar. 31, 2022).

[2] "Radiation Therapy for Cancer - NCI." https://www.cancer.gov/about-cancer/treatment/types/radiation-therapy (accessed Oct. 06, 2022).

[3] L. A. Bennie, H. O. McCarthy, and J. A. Coulter, "Enhanced nanoparticle delivery exploiting tumour-responsive formulations," *Cancer Nanotechnol*, vol. 9, no. 1, pp. 1–20, Nov. 2018, doi: 10.1186/S12645-018-0044-6/FIGURES/2.

[4] Y. Nakamura, A. Mochida, P. L. Choyke, and H. Kobayashi, "Nanodrug delivery: is the enhanced permeability and retention effect sufficient for curing cancer?," *Bioconjug Chem*, vol. 27, no. 10, pp. 2225–2238, Oct. 2016, doi: 10.1021/acs.bioconjchem.6b00437.

[5] R. Ngoune, A. Peters, D. von Elverfeldt, K. Winkler, and G. Pütz, "Accumulating nanoparticles by EPR: a route of no return," *J Control Release*, vol. 238, pp. 58–70, Sep. 2016, doi: 10.1016/j.jconrel.2016.07.028.

[6] U. Chitgupi, Y. Qin, and J. F. Lovell, "Targeted Nanomaterials for Phototherapy.," *Nanotheranostics*, vol. 1, no. 1, pp. 38–58, 2017, doi: 10.7150/ntno.17694.

[7] N. Rubio, A. Rajadurai, K. D. Held, K. M. Prise, H. L. Liber, and R. W. Redmond, "Real-time imaging of novel spatial and temporal responses to photodynamic stress.," *Free Radic Biol Med*, vol. 47, no. 3, pp. 283–90, Aug. 2009, doi: 10.1016/j.freeradbiomed.2009.04.024.

[8] H. Kolarova, P. Nevrelova, K. Tomankova, P. Kolar, R. Bajgar, and J. Mosinger, "Production of reactive oxygen species after photodynamic therapy by porphyrin sensitizers.," *Gen Physiol Biophys*, vol. 27, no. 2, pp. 101–5, Jun. 2008.

[9] P. Singh, S. Pandit, V. R. S. S. Mokkapati, A. Garg, V. Ravikumar, and I. Mijakovic, "Gold Nanoparticles in Diagnostics and Therapeutics for Human Cancer," *Int J Mol Sci*, vol. 19, no. 7, 2018, doi: 10.3390/ijms19071979.

[10] J. Berlanda, T. Kiesslich, V. Engelhardt, B. Krammer, and K. Plaetzer, "Comparative in vitro study on the characteristics of different photosensitizers employed in PDT," *J Photochem Photobiol B*, vol. 100, no. 3, pp. 173–180, 2010, doi: https://doi.org/10.1016/j.jphotobiol.2010.06.004.

[11] V. Amendola, R. Pilot, M. Frasconi, O. M. Maragò, and M. A. Iatì, "Surface plasmon resonance in gold nanoparticles: a review," *Journal of Physics: Condensed Matter*, vol. 29, no. 20, p. 203002, May 2017, doi: 10.1088/1361-648X/aa60f3.

[12] "SMOTE | Towards Data Science." https://towardsdatascience.com/smote-fdce2f605729 (accessed Oct. 12, 2022).

[13] "A Gentle Introduction to ImUnbalanced Classification." https://machinelearningmastery.com/what-is-imUnbalanced-classification/ (accessed Oct. 12, 2022).

[14] "sklearn.metrics.confusion_matrix — scikit-learn 1.1.2 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html (accessed Oct. 03, 2022).

[15] "ROC Curves & AUC: What Are ROC Curves | Built In." https://builtin.com/data-science/roc-curves-auc (accessed Oct. 03, 2022).