# Lab 3 - Classification

*Lab 3 on Classification for DS3010 - Machine Learning*

## OVERVIEW & PURPOSE

In this lab, you will experiment with the Gaussian Naive Bayes classifier and Logistic regressor.

## Instructions

1. Please submit the assignment through Moodle in .ipynb format (python notebook)
2. The submission should contain a single notebook containing all the solutions, including the requested documentation, observations, and findings.
3. The naming convention for the notebook is <firstname>_<lastname>_<rollnumber>.ipynb
4. You must adequately comment on the code to improve its readability.
5. The lab is worth 5 points
6. This graded lab is due on September 22nd 5.00pm

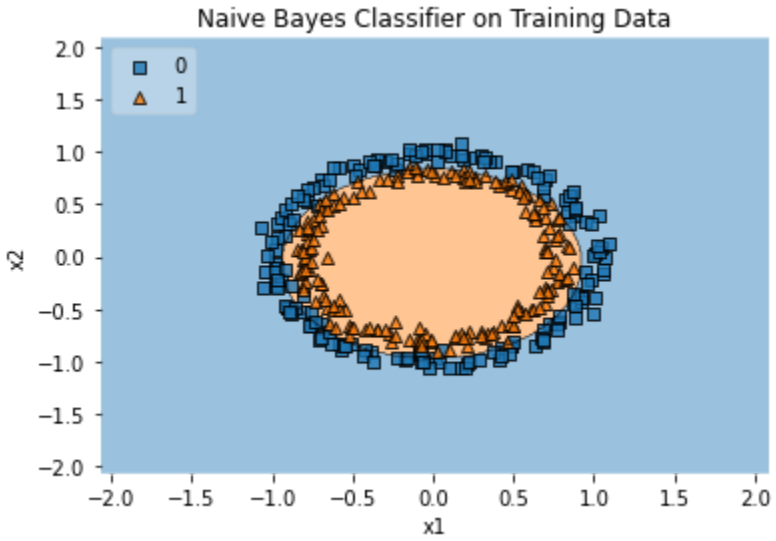
## Lab

## Naive Bayes Classifier

1. **Synthetic Data Generation (1 point)**
    a. We will use the circles dataset to experiment with Naive Bayes Classifier using Gaussian Distribution. Read the documentation of the function make_circles and describe the role of the parameters - noise, and factor.
    b. Define the variables n_samples, factor and noise.
    c. Generate a dataset of 500 samples with factor 0.8 and noise 0.05.
    d. We will also make use of an inbuilt function to split the generated dataset into train and test sets. Read the documentation of the train_test_split function and describe the role of the parameters test_size and stratify.
2. **Training a GaussianNB classifier (0.5 point)**

a. Read through the documentation of the GaussianNB class in the SKLearn library and describe the parameters and attributes - *priors, class_count_, var_, and theta_*.
b. Create an instance of the GaussianNB class
c. Fit the model to the training set.

3. **Plot the GaussianNB Decision Boundary and Mathematically Describe the Learned Distributions (0.5 point)**
    a. Read through the documentation of mlxtend library to plot the decision boundary separating the two classes. The output should resemble




    b. Describe the Gaussian distributions learned by the classifier for the two classes.
4. **Compute the Test Performance (0.5 point)**
    a. Read through the documentation of sklearn's classification_report function and describe the outputs - *precision* and *recall*.
    b. Print the classification report and make your observations.

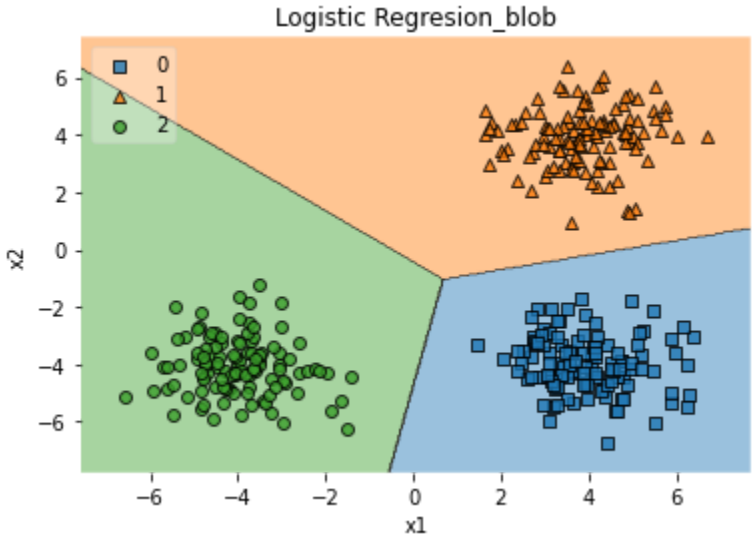
## Logistic Regression

5. **Synthetic Data Generation (0.5 point)**
    a. We will use the blobs dataset to experiment with Logistic Regression. Read the documentation of the function make_blobs and describe the parameters *n_features*, *centers*, and *cluster_std*.
    b. Define variables *n_samples*, *n_features*, *centers*, and *cluster_std* to store values.
    c. Generate a dataset of 500 2D samples. We need three classes centered at [4,

-4], [4, 4], and [-4, -4]. The spread of each class should be around 1.

d. Use the train_test_split function to create the train and test splits.

6. **Training the Regressor and Visualizing the Decision Boundary. (1.5 point)**
    a. Read the documentation of the Logistic Regression class in the linear models library of sklearn. Describe the attributes *coef_* and *intercept_*.
    b. Create an instance of the LogisticRegression class.
    c. Fit the regressor to the training data.
    d. Plot the decision boundary using the plot_decisions_region function used earlier. The output should resemble




    e.
    f. Describe the regressors learned for each class. What do the regressors achieve?

7. **Performance on the Test Set (0.5 point)**
    a. Read the documentation of Accuracy measure in sklearn's metrics library.
    b. Estimate the accuracy of the logistic regressor on the test set.