Lab 3
Linear Classification

Chien-Hui Su

Po-Chih Kuo

Introduction

In this lab, students will need to implement classification methods to accurately classify whether the patient has diabetes (1) or not (0) based on their Age, BMI, and Glucose.

Dataset

- Given 25000 records for training/validation, 5000 for testing
- The data used in this lab has 3 features of patients.
  - Age, BMI, and Glucose(血糖)
- Classes
  - 2 classes: no diabetes(0), diabetes(1)

Goal

- Predict if the patients have diabetes
- Implement the Perceptron
- Implement Fisher’s Linear Discriminant Analysis (LDA)
- Implement LDA classifier **using** Gaussian distributions and MAP estimation

Grading Policy

Item	Score
Part 1: Perceptron	35%
Part 2: Fisher’s LDA	35%
Part 3: LDA + MAP	25%
Report	5%

The Evaluation Metric

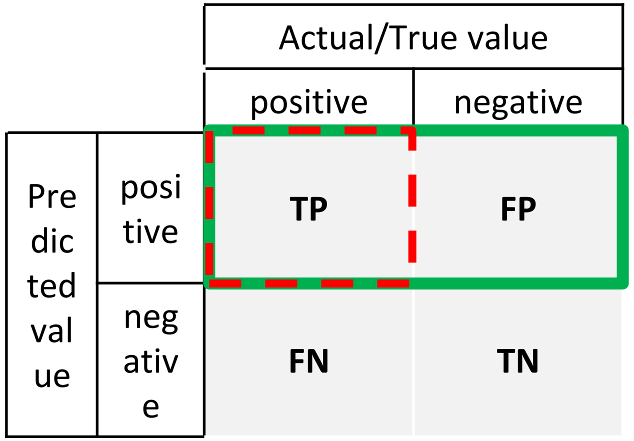
- F1-score

$$F1\text{-}score = 2 \times \frac{(\boldsymbol{Precision} \times \boldsymbol{Recall})}{(\boldsymbol{Precision}+\boldsymbol{Recall})}$$

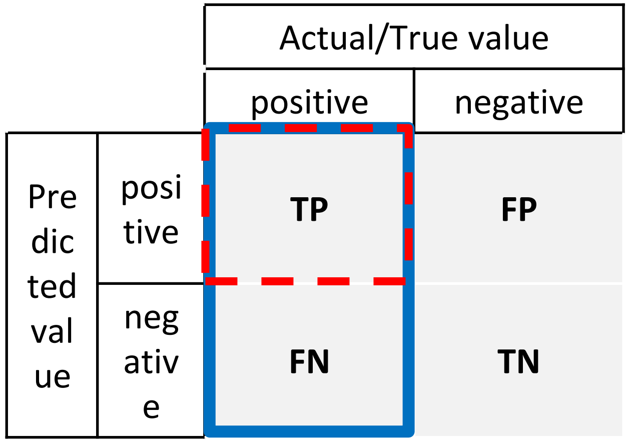
- For example
  - The class you predicted:

    $\hat{y}$ = [1, 1, 0, 0, 0, 0, 1]
  - Actual values:

    y = [0, 0, 0, 0, 0, 1, 1]
  - F1-score = 0.4



		Actual/True value	
		positive	negative
Predicted value	positive	**TP**	**FP**
	negative	**FN**	**TN**



		Actual/True value	
		positive	negative
Predicted value	positive	**TP**	**FP**
	negative	**FN**	**TN**

Grading Policy-Part 1: Perceptron (35%)

- Implement key functions of a Perceptron
- Submit the answer (.csv) to Kaggle **ML2024-Lab3-Perceptron**
- Get all if F1 score >= 0.5

Grading Policy-Part 2: LDA (35%)

- Implement key functions of Linear Discriminant Analysis (LDA)
- Submit the answer (.csv) to Kaggle **ML2024-Lab3-LDA**
- Get all if F1 score >= 0.6

Part 3 - LDA with MAP

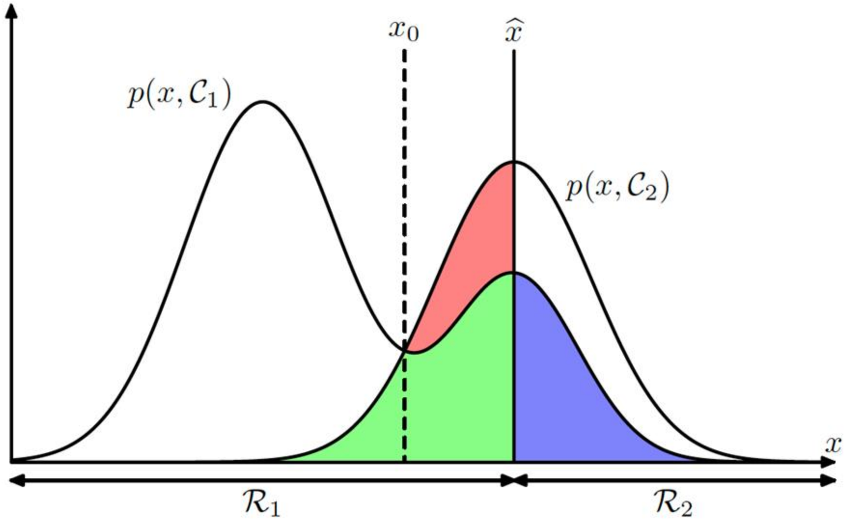


Figure 1.24 Schematic illustration of the joint probabilities $p(x, \mathcal{C}_k)$ for each of two classes plotted against x, together with the decision boundary $x = \widehat{x}$. Values of $x \geqslant \widehat{x}$ are classified as class $\mathcal{C}_2$ and hence belong to decision region $\mathcal{R}_2$, whereas points $x < \widehat{x}$ are classified as $\mathcal{C}_1$ and belong to $\mathcal{R}_1$. Errors arise from the blue, green, and red regions, so that for $x < \widehat{x}$ the errors are due to points from class $\mathcal{C}_2$ being misclassified as $\mathcal{C}_1$ (represented by the sum of the red and green regions), and conversely for points in the region $x \geqslant \widehat{x}$ the errors are due to points from class $\mathcal{C}_1$ being misclassified as $\mathcal{C}_2$ (represented by the blue region). As we vary the location $\widehat{x}$ of the decision boundary, the combined areas of the blue and green regions remains constant, whereas the size of the red region varies. The optimal choice for $\widehat{x}$ is where the curves for $p(x, \mathcal{C}_1)$ and $p(x, \mathcal{C}_2)$ cross, corresponding to $\widehat{x} = x_0$, because in this case the red region disappears. This is equivalent to the minimum misclassification rate decision rule, which assigns each value of x to the class having the higher posterior probability $p(\mathcal{C}_k|x)$.

(Chapter 4.5.1 of Pattern Recognition and Machine Learning. *Book by Christopher Bishop)*

Part 3 - LDA with MAP

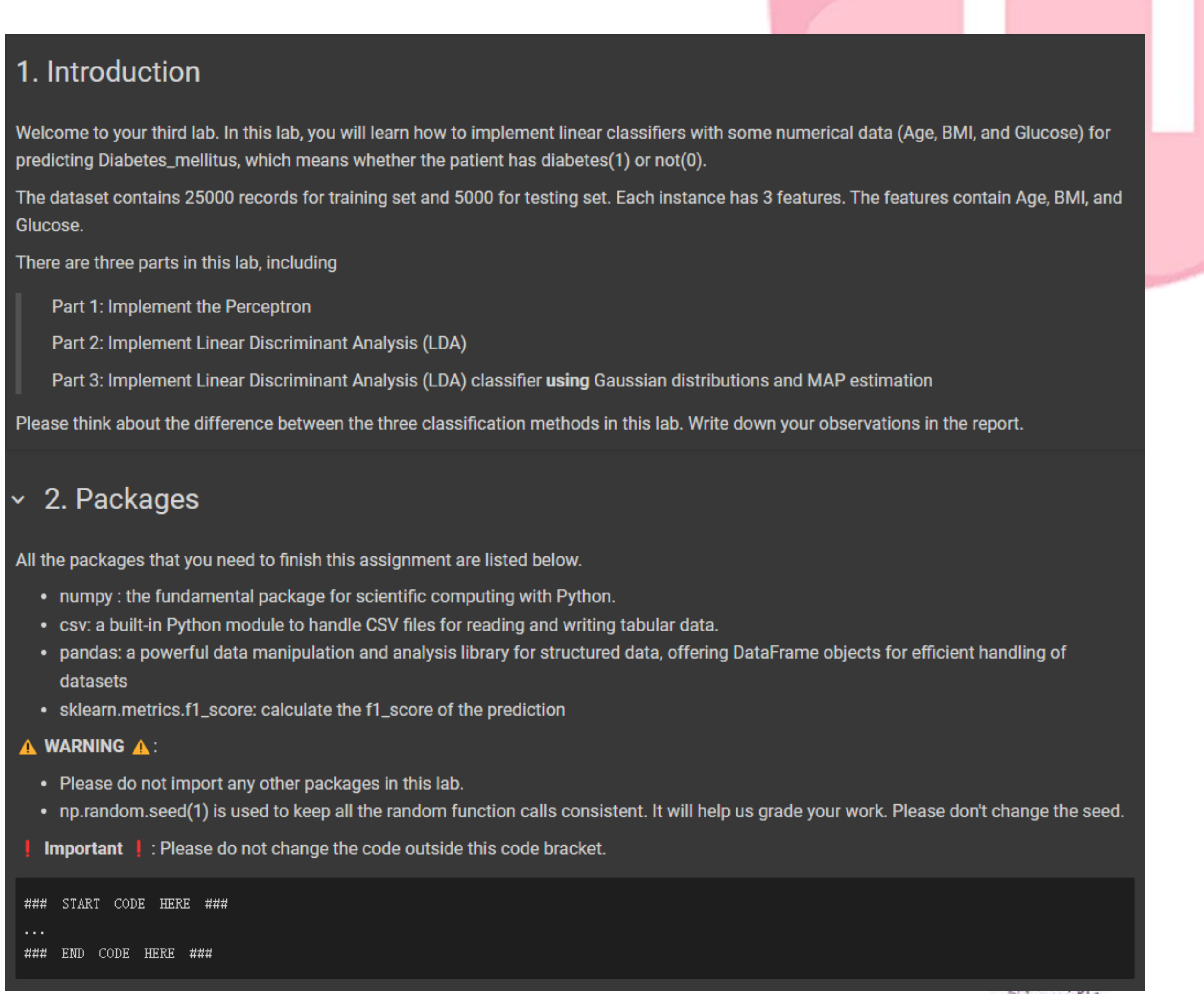
1. LDA projects the data onto a lower-dimensional space that maximizes class separability
2. After projection, we assume each class follows a Gaussian distribution in this new space. Computes the means, variances, and priors of each class in the LDA-projected space.
3. Implement the Gaussian density function.
4. Use MAP estimation
   a. For each test point, calculate its likelihood of belonging to each class using the likelihood function.
   b. Multiply these likelihoods by the class priors to get quantities proportional to the posterior probabilities.
   c. Predict based on the highest posterior probability.

Grading Policy-Part 3: LDA with MAP (25%)

- Implement key functions of Linear Discriminant Analysis (LDA) **using** Gaussian distributions and Maximum A Posterior (MAP) estimation.
- Submit the answer (.csv) to Kaggle **ML2024-Lab3-LDAMAP**
- Get all if F1 score >= 0.6

Template

- You must use the given file "Lab3_template.ipynb" to build the model
- Except for the imported packages in the template, you cannot use any other packages in this lab



1. Introduction

Welcome to your third lab. In this lab, you will learn how to implement linear classifiers with some numerical data (Age, BMI, and Glucose) for predicting Diabetes_mellitus, which means whether the patient has diabetes(1) or not(0).

The dataset contains 25000 records for training set and 5000 for testing set. Each instance has 3 features. The features contain Age, BMI, and Glucose.

There are three parts in this lab, including

> Part 1: Implement the Perceptron
>
> Part 2: Implement Linear Discriminant Analysis (LDA)
>
> Part 3: Implement Linear Discriminant Analysis (LDA) classifier **using** Gaussian distributions and MAP estimation

Please think about the difference between the three classification methods in this lab. Write down your observations in the report.

2. Packages

All the packages that you need to finish this assignment are listed below.

- numpy : the fundamental package for scientific computing with Python.
- csv: a built-in Python module to handle CSV files for reading and writing tabular data.
- pandas: a powerful data manipulation and analysis library for structured data, offering DataFrame objects for efficient handling of datasets
- sklearn.metrics.f1_score: calculate the f1_score of the prediction

⚠ **WARNING** ⚠:

- Please do not import any other packages in this lab.
- np.random.seed(1) is used to keep all the random function calls consistent. It will help us grade your work. Please don't change the seed.

❗ **Important** ❗ : Please do not change the code outside this code bracket.

```
### START CODE HERE ###
...
### END CODE HERE ###
```

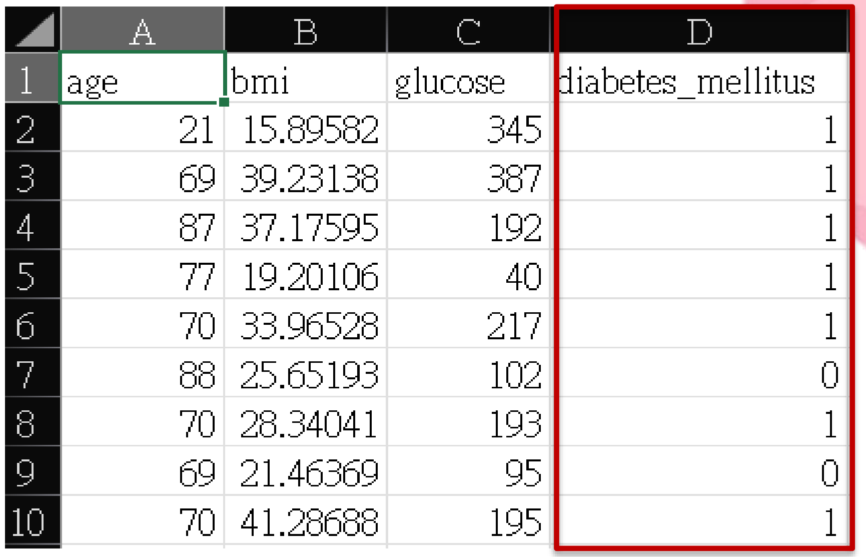
Input File Format

- There will be two input files:

1. “lab3_training.csv”
    - Label 0, 1
    - Each row has 3 features
    - Contains 25000 rows

2. “lab3_testing.csv”
    - Contains 5000 rows

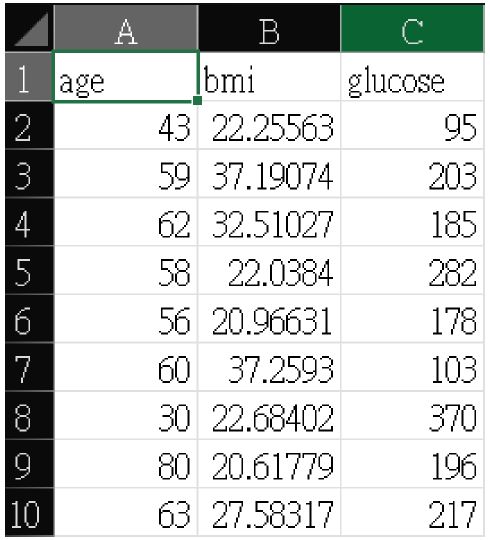
lab3_training.csv



	A	B	C	D
1	age	bmi	glucose	diabetes_mellitus
2	21	15.89582	345	1
3	69	39.23138	387	1
4	87	37.17595	192	1
5	77	19.20106	40	1
6	70	33.96528	217	1
7	88	25.65193	102	0
8	70	28.34041	193	1
9	69	21.46369	95	0
10	70	41.28688	195	1

label

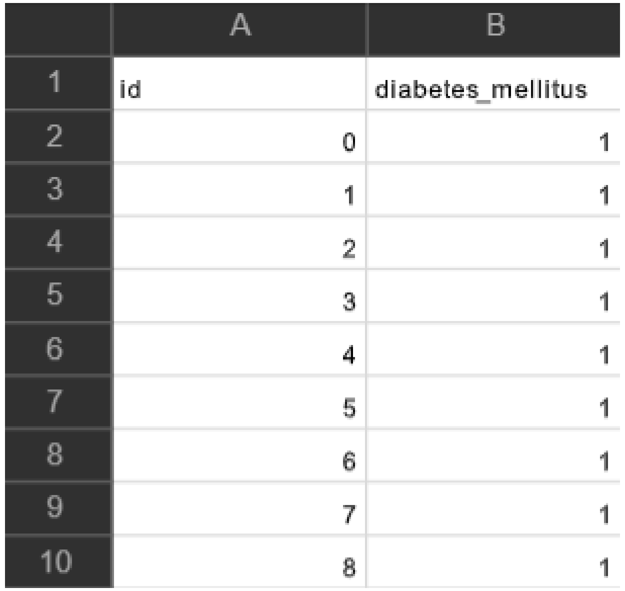
lab3_testing.csv



	A	B	C
1	age	bmi	glucose
2	43	22.25563	95
3	59	37.19074	203
4	62	32.51027	185
5	58	22.0384	282
6	56	20.96631	178
7	60	37.2593	103
8	30	22.68402	370
9	80	20.61779	196
10	63	27.58317	217

Output File Format

- There should be (5000+1) rows in your csv file
  - First row is the header [‘id’, ‘diabetes_mellitus’]
  - Your prediction answer should be either 0 or 1
  - Id starts from 0, and **diabetes_mellitus** is the predicted answer
- Please make sure that your output format is correct
- Submit the answer (.csv) to Kaggle **ML2024-Lab3-Perceptron, ML2024-Lab3-LDA, ML2024-Lab3-LDAMAP respectively**



	A	B
1	id	diabetes_mellitus
2	0	1
3	1	1
4	2	1
5	3	1
6	4	1
7	5	1
8	6	1
9	7	1
10	8	1

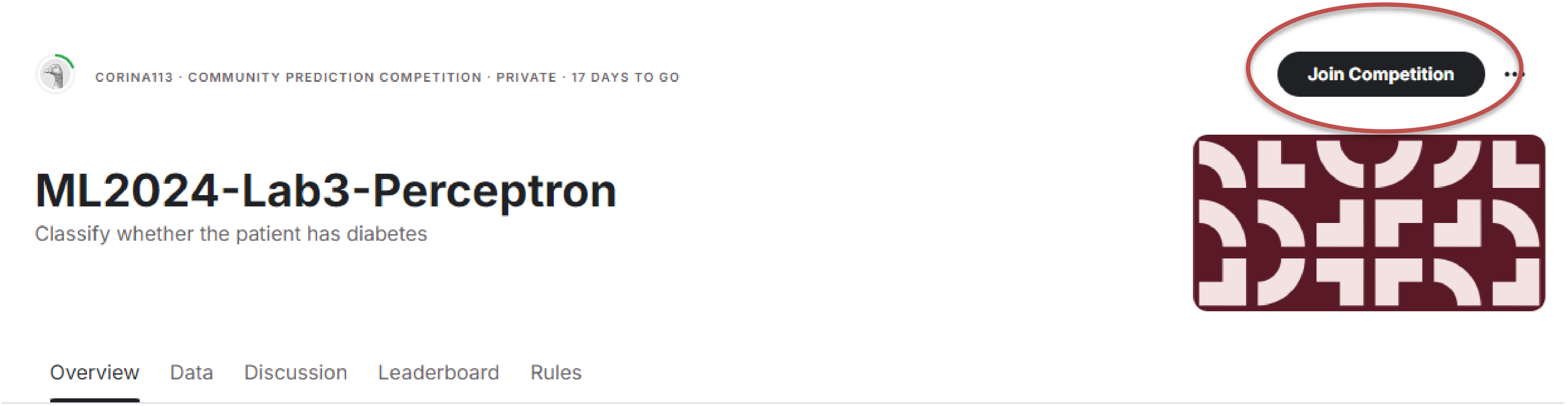
Kaggle

- We’ve created three competitions for each part respectively.
- Part 1 link: https://www.kaggle.com/t/45991d6a368e12568344cae35a1e4d9b
- Part 2 link: https://www.kaggle.com/t/2a4ed439f6bd997affa01473fd9f0b32
- Part 3 link: https://www.kaggle.com/t/6d62fd46fbb3421ea893761f77317c31

For each part of the lab, only a public score is provided. You can check if you pass the baseline directly.

Kaggle

- Please register your account.
- Click the ‘Join competition’ button to join.



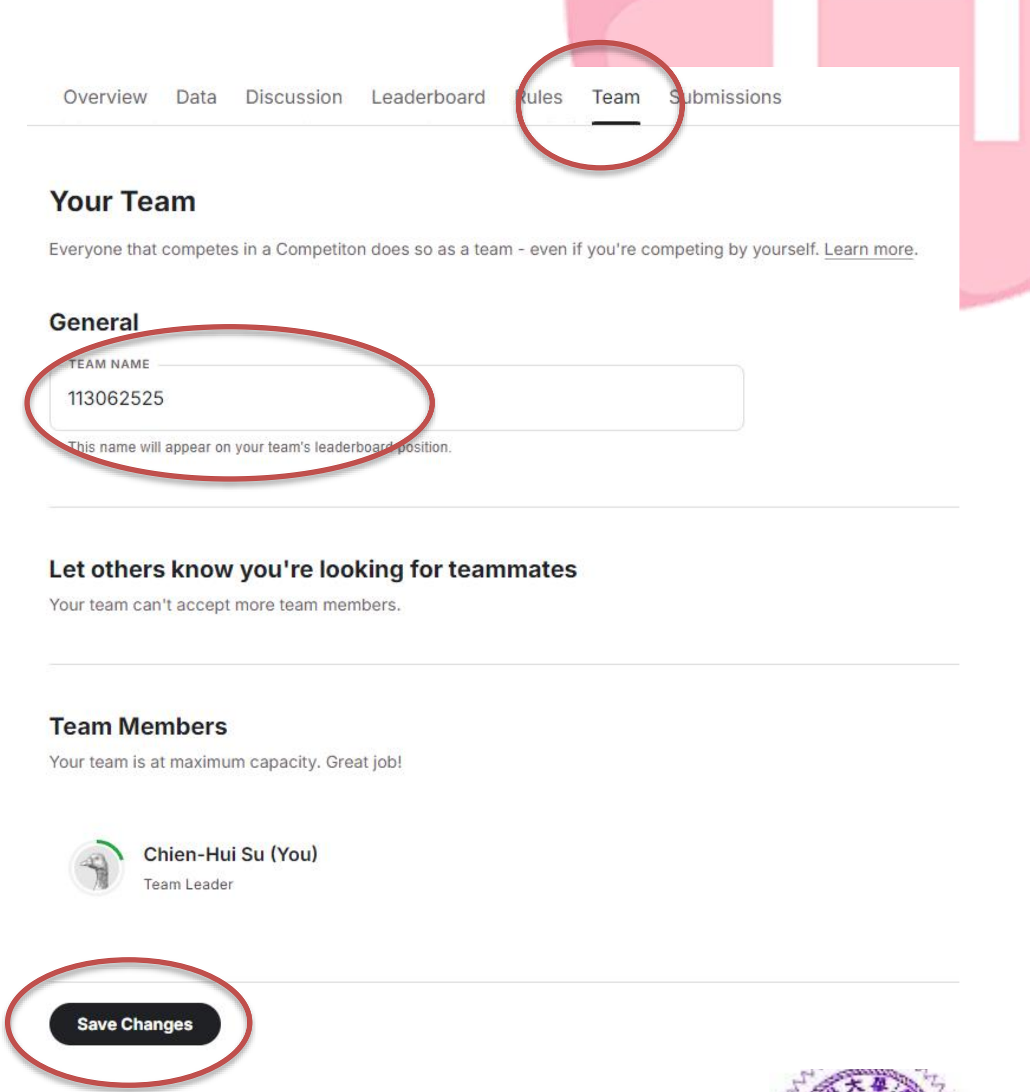
CORINA113 · COMMUNITY PREDICTION COMPETITION · PRIVATE · 17 DAYS TO GO

ML2024-Lab3-Perceptron

Classify whether the patient has diabetes

Kaggle

- After joining the competition, you should change your team name (each student is a team) to your **student ID**.
- **Please remember to SAVE CHANGES**
- You can submit 50 times per day.



Your Team

Everyone that competes in a Competiton does so as a team - even if you're competing by yourself. Learn more.

General

TEAM NAME

113062525

This name will appear on your team's leaderboard position.

Let others know you're looking for teammates

Your team can't accept more team members.

Team Members

Your team is at maximum capacity. Great job!

Chien-Hui Su (You)

Team Leader

Report

- Named as “**Lab3_report.pdf**”
- State the possible reason why the accuracy or F1-score change between Perceptron and LDA? (2%)
- Does MAP help? Why?(2%)
- Summarize how you solve the difficulty and your reflections (1%)
- No more than one page

Lab 3 Requirement

- Do it individually! Not as a team! (The team is for final project)
- Announce date: 2024/10/17
- Deadline: 2024/10/31 23:59 (Late submission is not allowed!)
- Submit the answers (csv) to corresponding Kaggle competition.
  - **ML2024-Lab3-Perceptron**
  - **ML2024-Lab3-LDA**
  - **ML2024-Lab3-LDAMAP**
- Hand in following files to eeclass in the following format (Do not compressed!)
  - Lab3.ipynb
  - Lab3_report.pdf
- Lab 3 would be covered on the exam next time.

Penalty

- 0 points if any of the following conditions happened
  - Plagiarism
  - Late submission
  - Not using a template or importing any other packages
  - No submission record on Kaggle (we cannot identify who you are)
  - Your submission was not generated by your code
- 5 Points would be deducted if your submission format is incorrect

Questions?

- TA: Chien-Hui Su (fabienne1023@gapp.nthu.edu.tw)
- Do not ask for debugging.