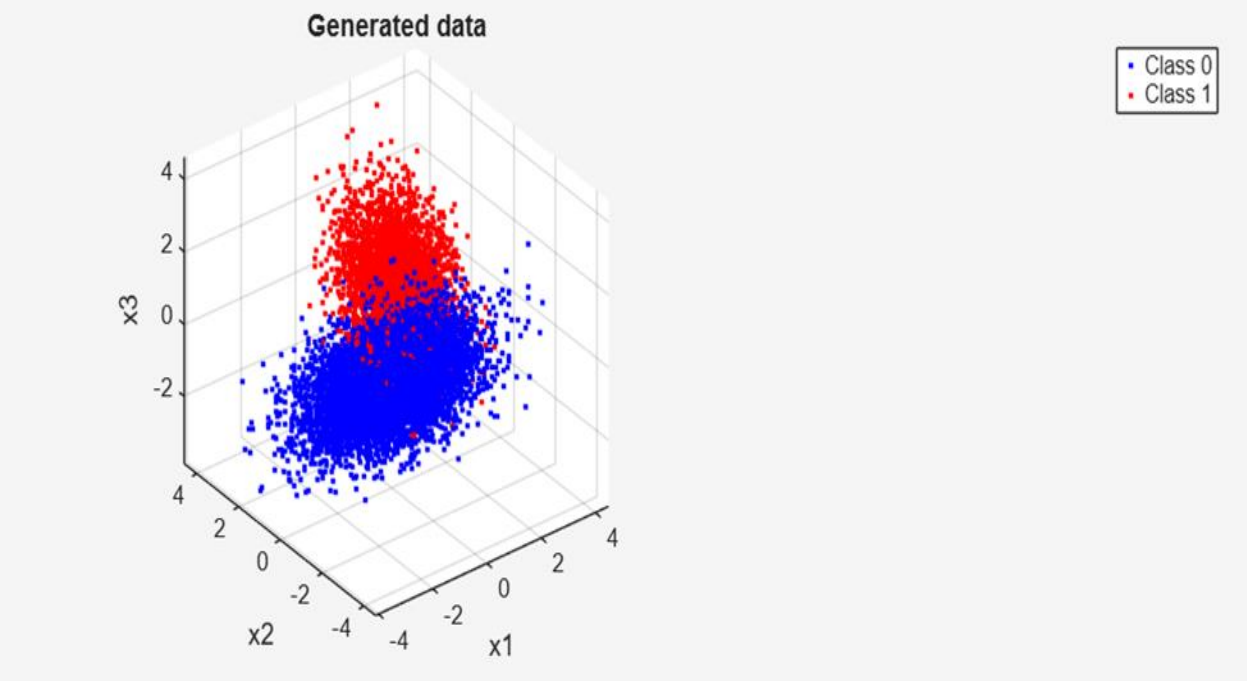
Assignment – 1

Prayag Sridhar
10/10/2025
Prof. Deniz Erdogmus

Question – 1: ERM classification using the knowledge of true pdf
3D scatter plot of the generated 10 000 Gaussian samples showing the separation between Class 0 (blue) and Class 1 (red) in the three-dimensional feature space.



Part-1.A: As per the minimum expected risk classification rule:

$$\frac{g(x|m_0, C_0)}{g(x|m_1, C_1)} \gtrless \frac{P(L=0)}{P(L=1)}\left(\frac{\lambda_{01} - \lambda_{00}}{\lambda_{10} - \lambda_{11}}\right) = \gamma$$

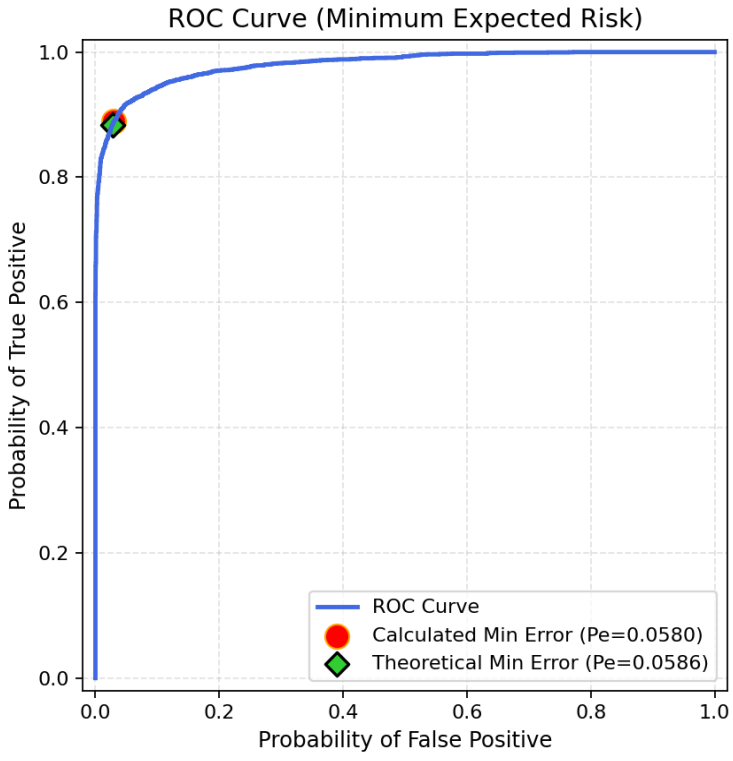
γ is the threshold function of class priors and fixed (non-negative) loss values.

λ_{ij} ranges from 0 to 1, Since we are to minimize the risk the values of λ_{01} and λ_{10} are set to 1 whereas the values of λ_{00} and λ_{11} are set to '0'.

$$\gamma = \frac{0.65}{0.35} \times \left(\frac{1-0}{1-0}\right) = 1.85714$$

$$\therefore \frac{P(L=1)}{P(L=0)} > 1.85714$$

Part-1.B: The figure below displays the ROC curve for minimum expected risk classification:



```
Minimum Expected Risk ROC values:
Calculated tau*: 0.503035,  gamma*: 1.653732
Calculated:    FPR=0.0299,  TPR=0.8899,  Pe=0.05799
Theoretical tau*: 0.619039,  gamma*: 1.857143
Theoretical:  FPR=0.0278,  TPR=0.8842,  Pe=0.05858
```

Part-1.C: The formula below is used to calculate the theoretical minimum risk

$P_e = 1- P(D = 0 \mid L = 0)*P(L=0) - P(D = 1 \mid L = 1)*P(L=1)$ or $P_e = FPR * p0 + (1 - TPR) * p1$

$p_0 = 0.65$; $p_1 = 0.35$

Calculated min-error probability: FPR = 0.0299; TPR = 0.8899; Tau = 0.503035; gamma = e^{tau} = 1.6537327408

$P_e = (0.0299)\ (0.65) + (1-0.8899)\ (0.35) = 0.019435 + 0.038535 = 0.05797$

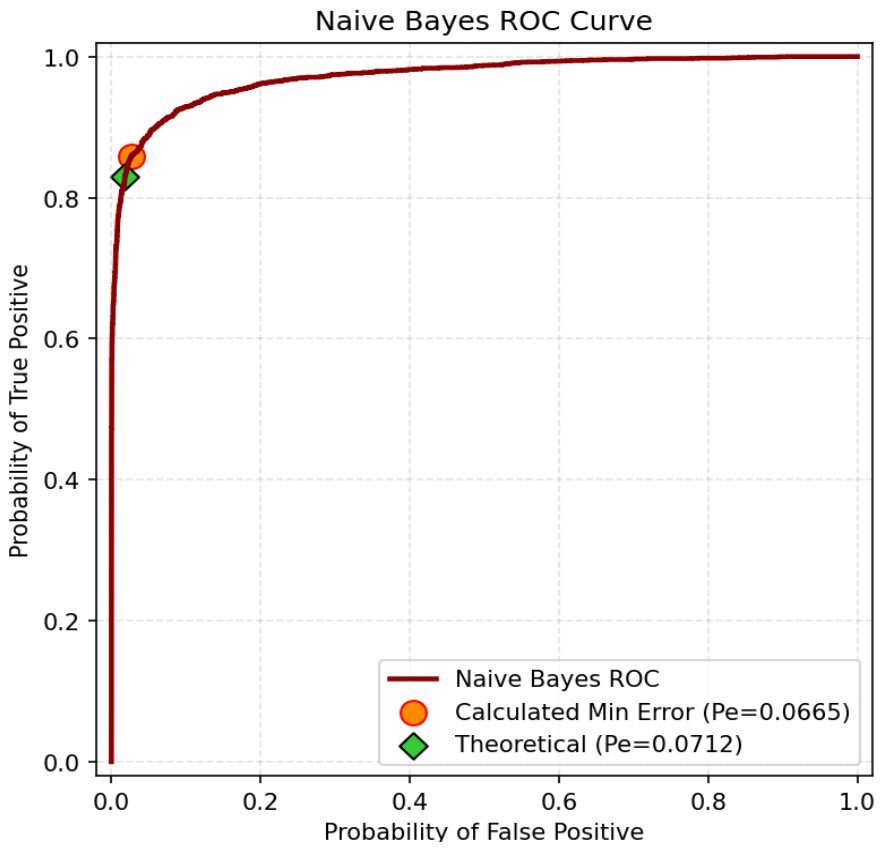
Theoretical probability: FPR = 0.0278; TPR = 0.8842; Tau = 0.619039; gamma = e^{tau} = 1.85714247

$P_e = (0.0278)\ (0.65) + (1-0.8842)\ (0.35) = 0.01807 + 0.04053 = 0.05860$

	Gamma	Min P_e
Calculated	1.6537327408	0.05797
Theoretical	1.85714247	0.05860

Part-B: Naive Bayes classification

The ROC curve of the naive bayes classification is given below:



```
Naive Bayes ROC calculations:
Calculated tau*: 0.252288; Calculated gamma*: 1.286967
Calculated: FPR=0.0264; Calculated TPR=0.8589; Calculated Pe=0.06654

Theoretical tau*: 0.619039; Theoretical gamma*: 1.857143
Theoretical: FPR=0.0180; Theoretical TPR=0.8301; Theoretical Pe=0.07119
```

The formula below is used to calculate the probability:
$P_e = 1- P\ (D = 0 \mid L = 0)*P(L=0) - P\ (D = 1 \mid L = 1)*P(L=1)$ or P_e = FPR * p0 + (1 - TPR) * p1
$p_0 = 0.65$; $p_1 = 0.35$
Calculated min-error probability: FPR = 0.0264; TPR = 0.8589; Tau = 0.252288; gamma = e^{tau} = 1.2869663
$P_e = (0.0264)\ (0.65) + (1-0.8589)\ (0.35) = 0.066545$

Theoretical probability: FPR = 0.0278; TPR = 0.8842; Tau = 0.619039; gamma = e^{tau} = 1.85714247
$P_e = (0.0180)\ (0.65) + (1-0.8301)\ (0.35) = 0.071165$

	Gamma	Minimum P_e
Calculated	1.28696663	0.066545
Theoretical	1.85714247	0.071165

Did this model mismatch negatively impact your ROC curve and minimum achievable probability of error?

When the Naive Bayes assumption was applied by setting both covariance matrices to the identity, the classifier performance declined compared to the Bayes-optimal case in Part A. Because this model ignores the correlations that actually exist among the three features, the likelihood estimates no longer represent the true class-conditional densities. As a result, the ROC curve flattened and shifted closer to the diagonal, indicating poorer discrimination between the two classes. The minimum probability of error also increased slightly from roughly 0.058 in the true model to about 0.067 in this current model showing that the classifier made more wrong decisions overall. In short, the independence assumption simplified the computation but caused a noticeable loss of accuracy, confirming that this model mismatch negatively affected both the ROC curve and the minimum achievable probability of error.

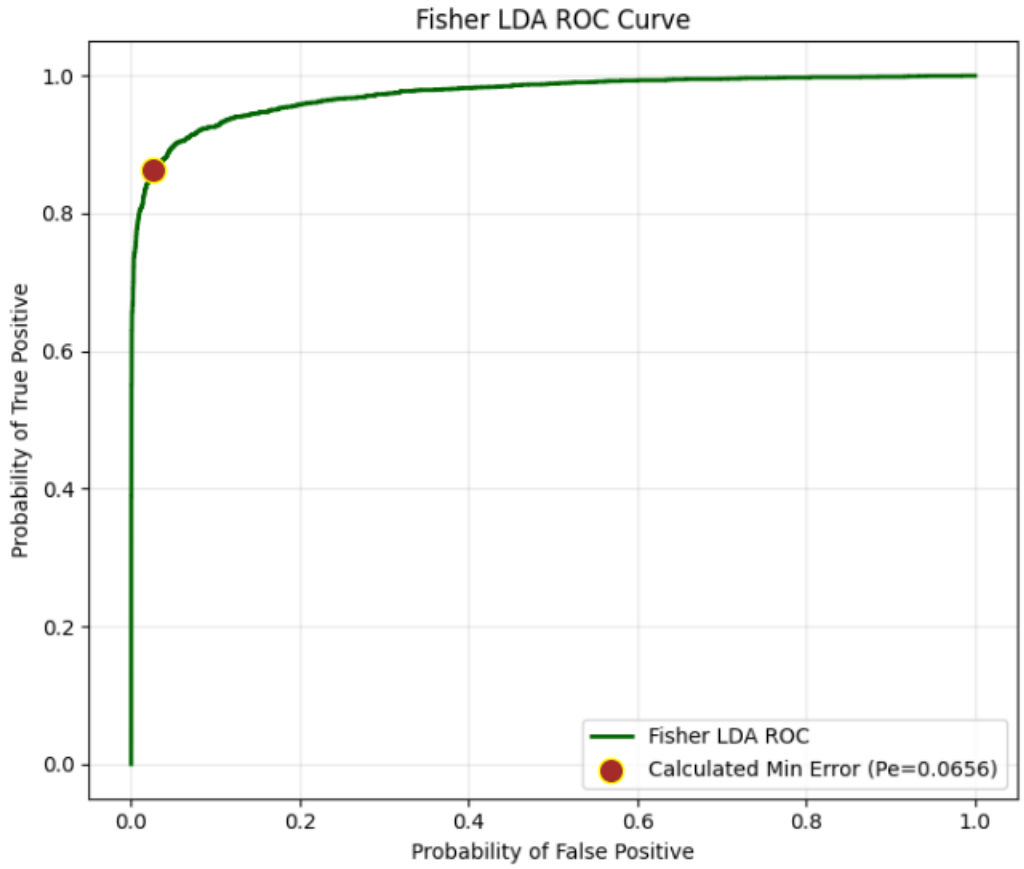
Part-C: Fisher LDA classifier

The Fisher LDA projection weight vector W_{LDA}:

```
Within-class scatter (S_W):
 [[ 2.02347093 -0.20561545  0.13165125]
 [-0.20561545  2.04046348 -0.19696772]
 [ 0.13165125 -0.19696772  2.04823671]]

Fisher LDA projection vector (w_LDA): [0.77549539 0.89368419 0.77131811]
```

ROC curve:




```
Fisher LDA calculations:
Threshold (tau): 0.6492
FPR: 0.0268
TPR: 0.8624
Min Pe: 0.06558
```

Performance comparison of Fisher LDA classifier to the previous 2 classifiers:

The Fisher LDA classifier's performance is in between both of the Bayes-optimal and Naive Bayes findings. The fisher LDA takes sample estimates of the class means and covariances and assumes that both classes have the same dispersion within the class. This means that it captures part of the correlation structure that the Naive Bayes model missed.

The ROC curve for LDA stays near to the upper-left corner, which means that it can tell the two classes apart very well, but not as well as the Bayes-optimal curve from Part A. The error probability for LDA (0.06558) is a little higher than the true-pdf case (about 0.0580) but better than the Naive Bayes case (about 0.0665). In general, LDA strikes a solid balance: its linear projection

gives good separation at a low computational cost, resulting in accuracy that is almost perfect while avoiding the bigger drop in performance that comes from the independence requirement in the Naive Bayes classifier.

Question-2: Gaussian class conditional pdf

Part-A:
Here I considered a two-dimensional random vector $X = [x_1, x_2]^T$ belonging to one of four possible classes $L \in \{1,2,3,4\}$.

Each class is modeled by a Gaussian class-conditional pdf $p(x \mid L = j) = N(x; MU_j, SIGMA_j)$ with equal priors $P(L=j) = 0.25$

The mean and covariance matrices taken are:
Mean of class 1: $[2, 1]^T$
Mean of class 2: $[0, -3]^T$
Mean of class 3: $[-2, 1]^T$
Mean of class 4: $[0, 0]^T$

Covariance of class 1: [4 -3;-3 4]
Covariance of class 2: [3 0;0 1]
Covariance of class 3: [4 3;3 4]
Covariance of class 4: [1 0;0 1]

Decision rule (MAP): Under 0-1 loss, the classifier that minimizes probability of error chooses the class with the highest posterior probability:
$D(x)=\text{argmax}\ [p(x|L=i)P(L=i)]$

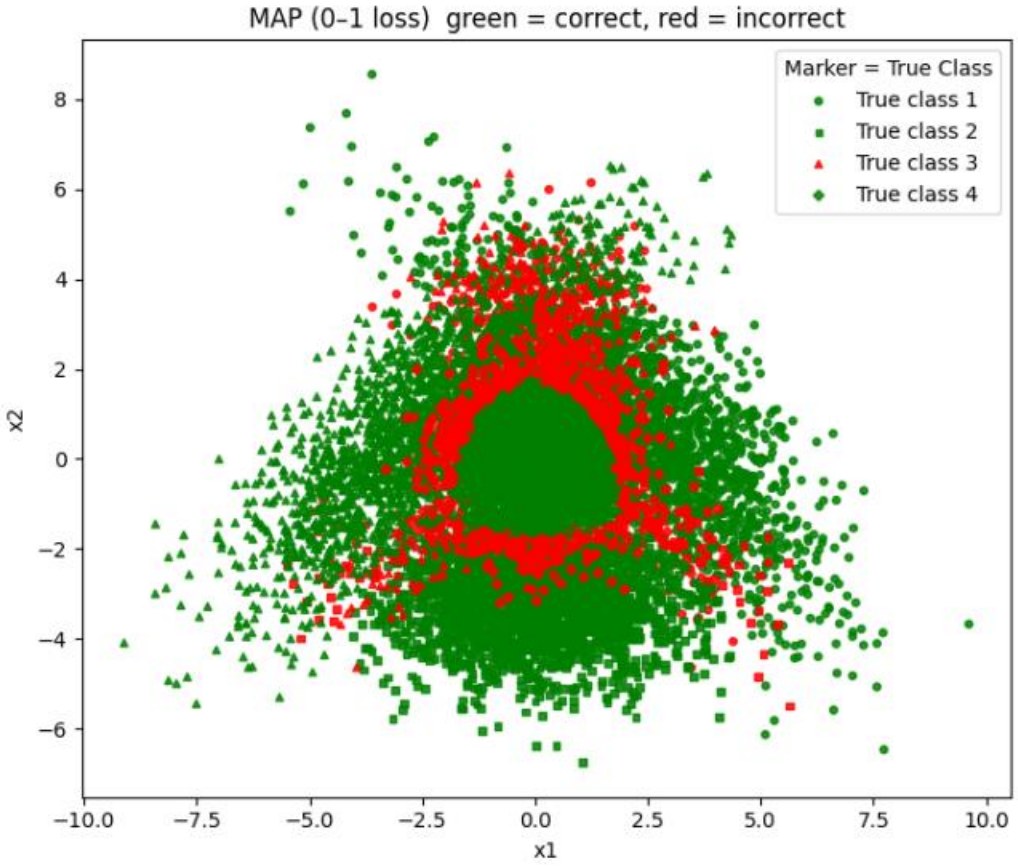
Since all priors are equal, this simplifies to picking the class with the largest likelihood $p(x \mid L = i)$.

The confusion matrix from the 10000 samples:

Confusion Matrix for MAP:

	L=1	L=2	L=3	L=4
D=1	0.765	0.030	0.102	0.077
D=2	0.019	0.899	0.021	0.059
D=3	0.120	0.022	0.774	0.076
D=4	0.096	0.049	0.102	0.788

Visualization:
The samples were drawn in 2-D scatter plot where green implies correct and red implies incorrect.



Part-B: Expected Risk Minimization

The loss matrix used was:
$\Lambda = [0\ 10\ 10\ 100;\ 1\ 0\ 10\ 100;1\ 1\ 0\ 100;1\ 1\ 1\ 0]$

Decision rule: The expected-risk classifier chooses the class that minimizes conditional risk:

$R_i(x) = \text{Sigma}\ \Lambda_{ij}\ P(L=j \mid x)$ and decides $D(x) = \arg\min R_i(x)$.

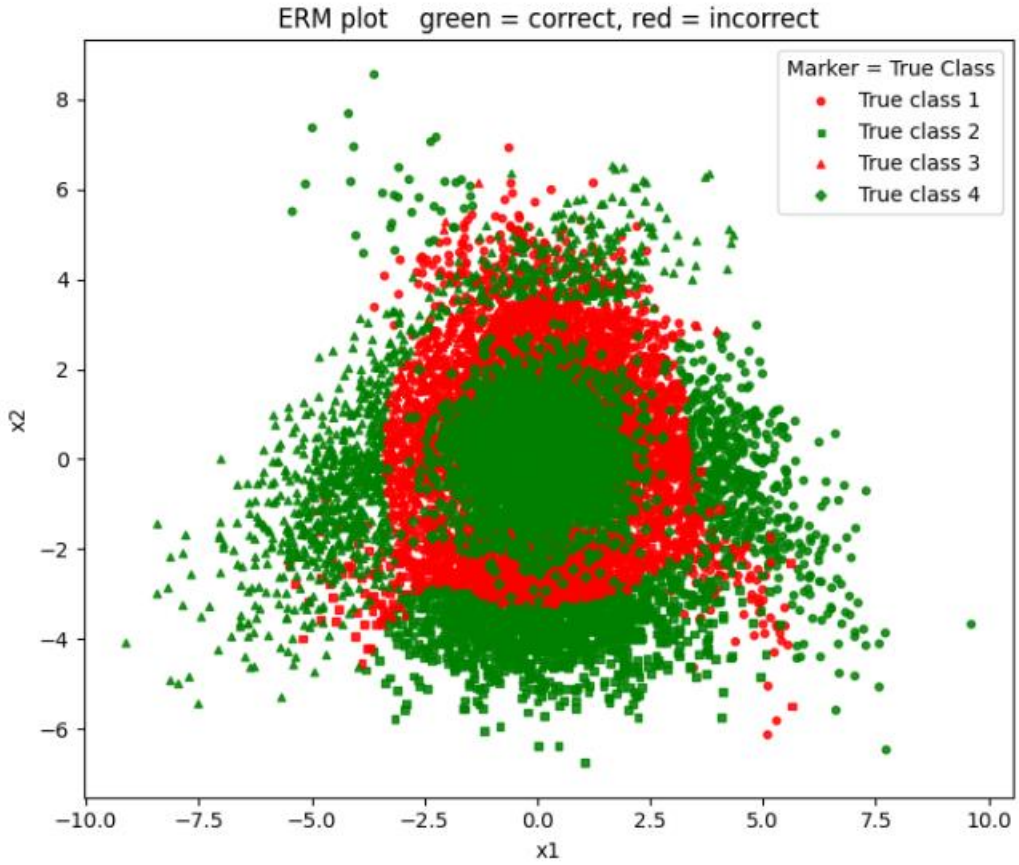
Confusion Matrix for ERM:

	L=1	L=2	L=3	L=4
D=1	0.219	0.004	0.002	0.000
D=2	0.039	0.492	0.002	0.001
D=3	0.102	0.026	0.384	0.000
D=4	0.640	0.478	0.613	0.998

Estimated Minimum Expected Risk: 0.5309

The estimated minimum risk from the 10000 samples is: 0.5309

Visualization:
A similar scatter plot was made for the ERM case.



The ERM classifier clearly leans toward classifying ambiguous points as class 4, which makes sense given the high penalty (100) for mistakes on that class.

Question – 3:
In this experiment, I implemented a minimum-probability-of-error classifier under the assumption that the class-conditional distributions of the observed features are multivariate Gaussian. Two datasets from the UCI Machine Learning Repository were analyzed:

1. Wine Quality (White) – 4898 samples, 11 physicochemical features, and discrete quality scores ranging from 0 to 10.
2. Human Activity Recognition (HAR) – 10299 samples, 561 features, and 6 activity labels collected from smartphone sensors.

The goal is to estimate class priors, means, and covariance matrices from data, apply a Bayes decision rule that minimizes the expected classification error, and evaluate the resulting confusion matrices. I also examined whether the Gaussian assumption is reasonable for these datasets.

Modeling Assumptions:
Each class L = k is modeled as a multivariate Gaussian: $p(x \mid L = k) = N(x \mid Mu_k, C_k)$.

Class priors were estimated using relative sample frequency: $\hat{Pi_k} = N_k / N$ where N_k and N are the number of samples in class k and the total number of samples respectively.

Sample means and covariances were calculated as:

$$\hat{Pi}_k = (1 / N_k) \Sigma x_i,$$
$$\hat{C}_k = (1 / (N_k - 1)) \Sigma (x_i - \hat{\mu_k})(x_i - \hat{\mu_k})^T.$$

Because covariance matrices can be ill-conditioned in high-dimensional settings, a regularization term was added: $C_k(reg) = \hat{C}_k + \lambda_k I$; $\lambda_k = alpha \cdot (\text{trace}(\hat{C}_k) / \text{rank}(\hat{C}_k))$ where *alpha* = 0.05 ensures positive definite covariance matrices without over-smoothing.

Classification Rule:
For each sample x, the log-discriminant function is computed for each class:
$g_k(x) = \log \hat{Pi}_k - 0.5 \log|C_k(reg)| - 0.5 (x - \hat{Mu}_k)^T (C_k(reg))^{-1} (x - \hat{Mu}_k)$.

The predicted label is $\hat{L} = \arg\max_k g_k(x)$.

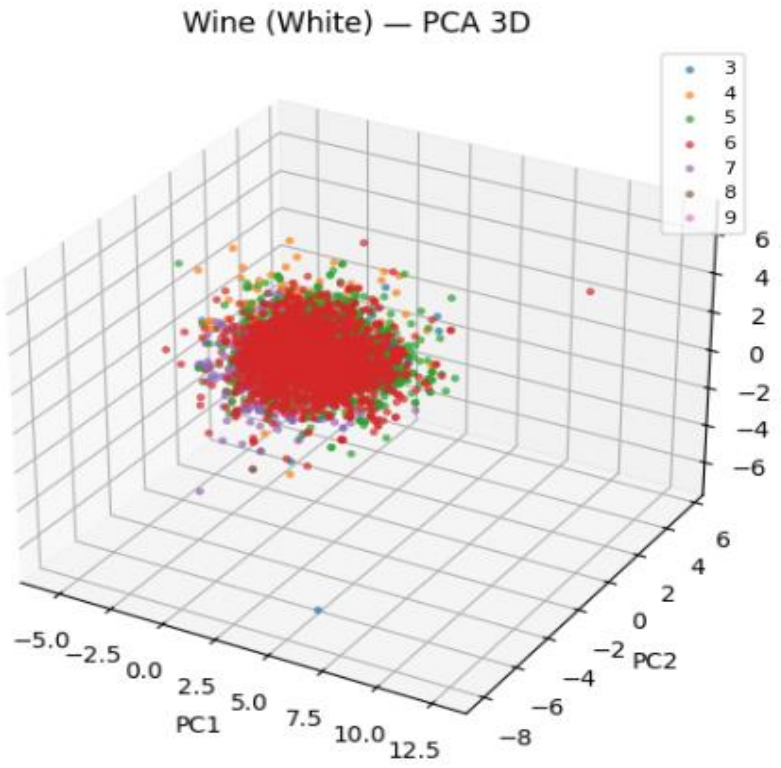
For both datasets, class priors were derived directly from the empirical proportions of samples in each category.

Results:
1: Wine Quality (White)

```
Wine (White): 4898 samples, 11 features, classes=[3 4 5 6 7 8 9]

Parameter summary per class:
class 3: n=20, prior=0.0041, lambda=74.8361, min_eig(C_reg)=74.8361
class 4: n=163, prior=0.0333, lambda=14.629, min_eig(C_reg)=14.629
class 5: n=1457, prior=0.2975, lambda=10.467, min_eig(C_reg)=10.467
class 6: n=2198, prior=0.4488, lambda=9.00425, min_eig(C_reg)=9.00425
class 7: n=880, prior=0.1797, lambda=5.76446, min_eig(C_reg)=5.76446
class 8: n=175, prior=0.0357, lambda=6.23876, min_eig(C_reg)=6.23876
class 9: n=5, prior=0.0010, lambda=7.36814, min_eig(C_reg)=7.36814

Wine (White): Confusion Matrix (rows=true, cols=pred)
    P3  P4  P5   P6    P7  P8  P9
T3   2   1   0    6     9   2   0
T4   1   0   2   42   117   1   0
T5   1   1  33  423   994   5   0
T6   1   0  25  426  1741   5   0
T7   0   0   0   64   811   5   0
T8   0   0   0   10   160   5   0
T9   0   0   0    0     5   0   0
Wine (White): Estimated P(error) = 0.7393
class 3: error = 0.9000
class 4: error = 1.0000
class 5: error = 0.9774
class 6: error = 0.8062
class 7: error = 0.0784
class 8: error = 0.9714
class 9: error = 1.0000
```



The overall training-set error rate was = 0.7393, consistent with the moderate separability observed in the 3-D PCA plot, where points formed a large, continuous cluster rather than distinct groups.

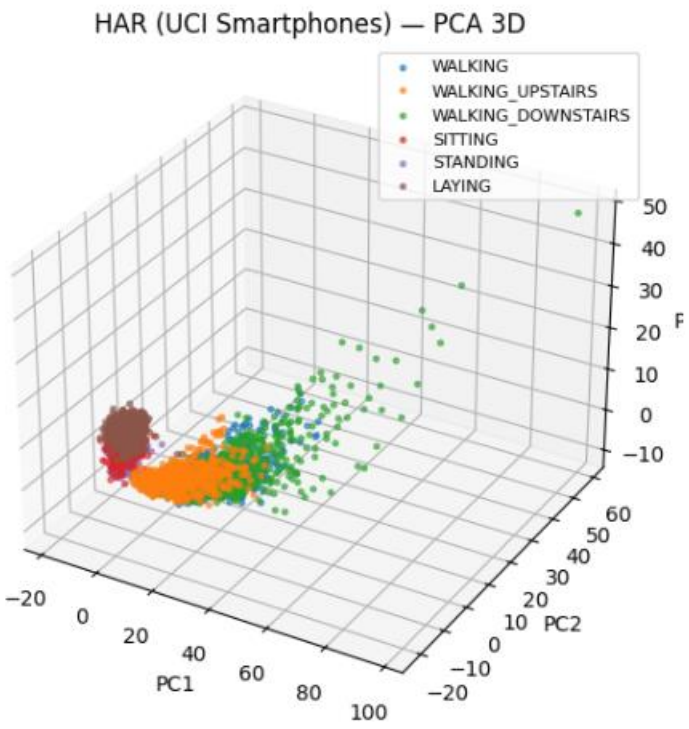
2: Human Activity Recognition (HAR)

For the HAR dataset, the classifier was trained using 10,299 samples spanning six activity categories. Regularization values λ_k were much smaller (like 10^{-2}) since the high-dimensional sensor data already exhibited well-spread eigenvalues.

```
HAR (UCI Smartphones): 10299 samples, 561 features, classes=[1 2 3 4 5 6]

Parameter summary per class:
class 1: n=1722, prior=0.1672, lambda=0.00214862, min_eig(C_reg)=0.00214862
class 2: n=1544, prior=0.1499, lambda=0.00220264, min_eig(C_reg)=0.00220264
class 3: n=1406, prior=0.1365, lambda=0.00302286, min_eig(C_reg)=0.00302286
class 4: n=1777, prior=0.1725, lambda=0.00171383, min_eig(C_reg)=0.00171383
class 5: n=1906, prior=0.1851, lambda=0.00154645, min_eig(C_reg)=0.00154645
class 6: n=1944, prior=0.1888, lambda=0.00214628, min_eig(C_reg)=0.00214628

HAR (UCI Smartphones): Confusion Matrix (rows=true, cols=pred)
      P1    P2    P3    P4    P5    P6
T1  1720     1     1     0     0     0
T2     0  1544     0     0     0     0
T3     5   121  1280     0     0     0
T4     0     0     0  1198   579     0
T5     0     0     0     0  1906     0
T6     0     0     0     0     0  1944
HAR (UCI Smartphones): Estimated P(error) = 0.0686
class 1: error = 0.0012
class 2: error = 0.0000
class 3: error = 0.0896
class 4: error = 0.3258
class 5: error = 0.0000
class 6: error = 0.0000
```



The train-set confusion matrix displayed strong diagonal dominance with overall accuracy = 93.14 % (P(error) = 0.0686).

The class-wise errors were lowest for WALKING, WALKING_UPSTAIRS, and WALKING_DOWNSTAIRS, while SITTING and STANDING occasionally overlapped; these postures produce similar accelerometer and gyroscope readings.

The 3-D PCA visualization further confirmed this: dynamic activities formed well-separated elongated manifolds, whereas static activities clustered closely together.

Online repository where I have uploaded all the codes:
https://github.com/PRAYAG2000n/Intro_to_ML_EECE5644/tree/main/Assignment%20-%201

Citations and References:
LDA: https://github.com/rajs96/QDA-LDA-Classifier
https://github.com/utsavberi/MachineLearning_ClassificationAndRegression
Gaussian Mixture Model From Scratch: https://github.com/DandiMahendris/Gaussian-mixture-from-scratch
Naïve Bayes: https://github.com/gbroques/naive-bayes
https://kuleshov-group.github.io/aml-book/contents/lecture7-gaussian-discriminant-analysis.html