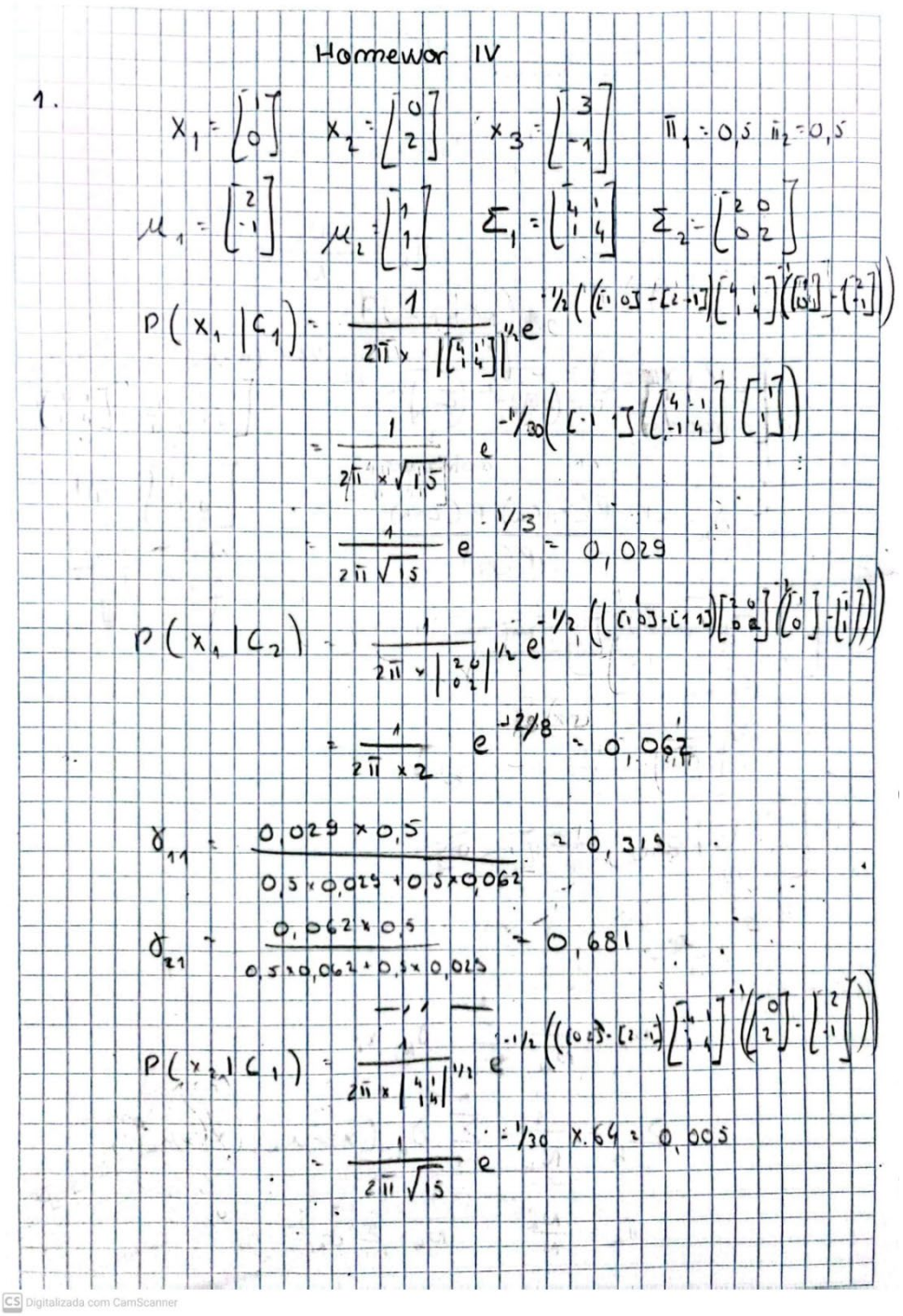
Aprendizagem 2024/25

Homework IV – Group 066

(ist106794, ist1107301)

I. Pen-and-paper

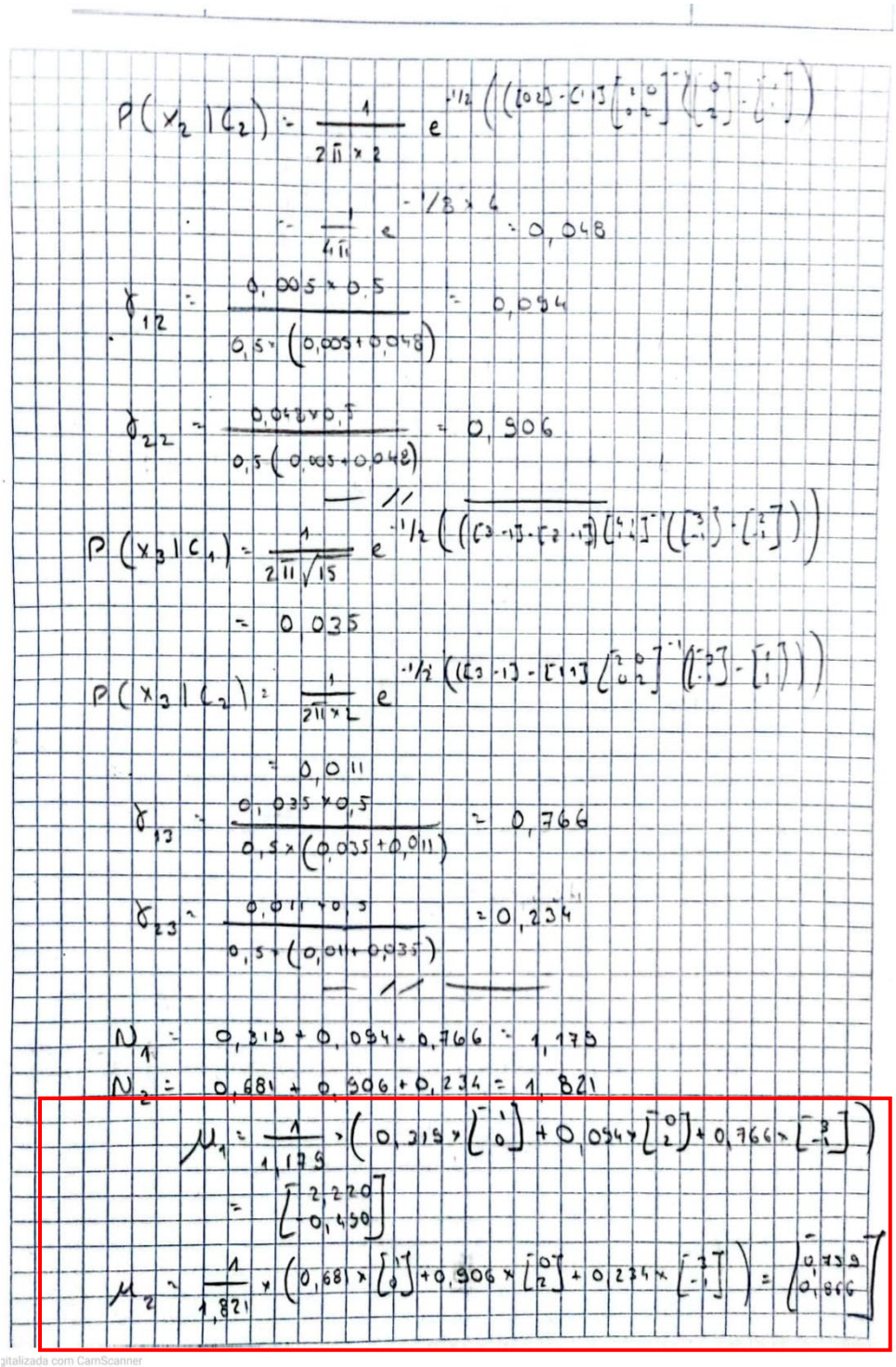
1)



Aprendizagem 2024/25

Homework IV – Group 066

(ist106794, ist1107301)



$$P(x_2|C_2) = \frac{1}{2\pi \times 2} e^{-1/2\left(\left([0\ 2]-[1\ 1]\right)\begin{bmatrix}1 & 0\\ 0 & 2\end{bmatrix}^{-1}\left(\begin{bmatrix}0\\2\end{bmatrix}-\begin{bmatrix}1\\1\end{bmatrix}\right)\right)}$$

$$= \frac{1}{4\pi} e^{-1/8 \times 4} = 0,048$$

$$\gamma_{12} = \frac{0,005 \times 0,5}{0,5 \times (0,005 + 0,048)} = 0,094$$

$$\gamma_{22} = \frac{0,048 \times 0,5}{0,5 \times (0,005 + 0,048)} = 0,906$$

$$P(x_3|C_1) = \frac{1}{2\pi\sqrt{15}} e^{-1/2\left(\left([3\ -1]-[2\ -1]\right)\begin{bmatrix}4 & 1\\ 1 & 4\end{bmatrix}^{-1}\left(\begin{bmatrix}3\\-1\end{bmatrix}-\begin{bmatrix}2\\1\end{bmatrix}\right)\right)}$$

$$= 0,035$$

$$P(x_3|C_2) = \frac{1}{2\pi \times 2} e^{-1/2\left(\left([3\ -1]-[1\ 1]\right)\begin{bmatrix}1 & 0\\ 0 & 2\end{bmatrix}^{-1}\left(\begin{bmatrix}3\\-1\end{bmatrix}-\begin{bmatrix}1\\1\end{bmatrix}\right)\right)}$$

$$= 0,011$$

$$\gamma_{13} = \frac{0,035 \times 0,5}{0,5 \times (0,035 + 0,011)} = 0,766$$

$$\gamma_{23} = \frac{0,011 \times 0,5}{0,5 \times (0,011 + 0,035)} = 0,234$$

$$N_1 = 0,315 + 0,094 + 0,766 = 1,175$$

$$N_2 = 0,681 + 0,906 + 0,234 = 1,821$$

$$\mu_1 = \frac{1}{1,175} \times \left(0,315 \times \begin{bmatrix}-1\\0\end{bmatrix} + 0,094 \times \begin{bmatrix}0\\2\end{bmatrix} + 0,766 \times \begin{bmatrix}-3\\2\end{bmatrix}\right)$$

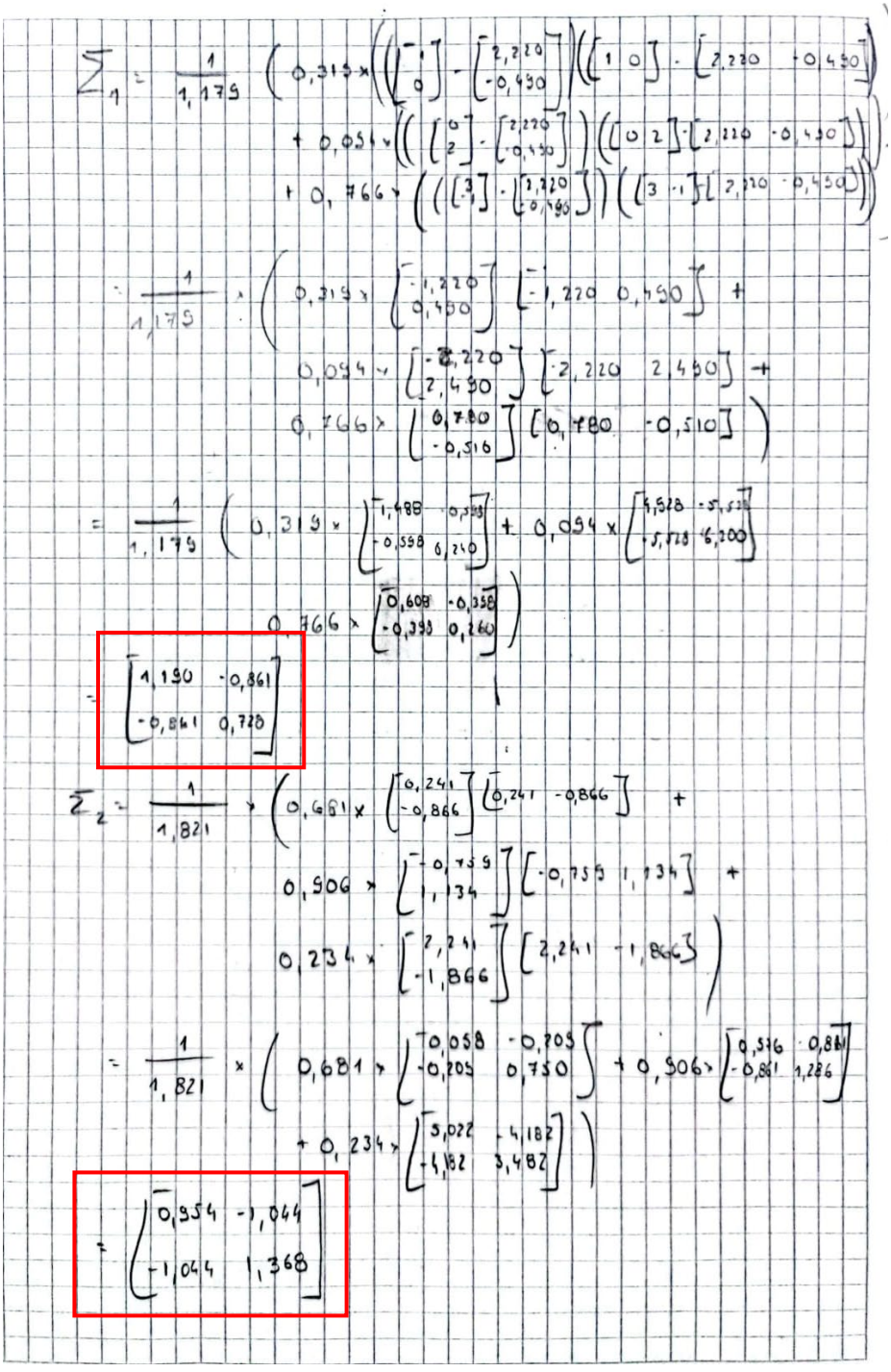
$$= \begin{bmatrix}-2,220\\0,450\end{bmatrix}$$

$$\mu_2 = \frac{1}{1,821} \times \left(0,681 \times \begin{bmatrix}1\\0\end{bmatrix} + 0,906 \times \begin{bmatrix}0\\2\end{bmatrix} + 0,234 \times \begin{bmatrix}-3\\-1\end{bmatrix}\right) = \begin{bmatrix}0,739\\0,866\end{bmatrix}$$

Aprendizagem 2024/25

Homework IV – Group 066

(ist106794, ist1107301)



TÉCNICO LISBOA

Aprendizagem 2024/25

Homework IV – Group 066

(ist106794, ist1107301)

$\pi_1 = \frac{1,179}{1,179+1,821} = 0,393$

$\pi_2 = \frac{1,821}{1,821+1,179} = 0,607$

2nd Epoch

$P(x_1|C_1) = \frac{1}{2\pi \times \sqrt{0,125}} e^{-\frac{1}{2}\left(([1\ 0]-[2,22\ \ -0,45])\begin{bmatrix}1,150 & -0,861\\ -0,861 & 0,728\end{bmatrix}^{-1}\left(\begin{bmatrix}1\\0\end{bmatrix}-\begin{bmatrix}2,2\\-0,\end{bmatrix}\right)\right.}$

$= 0,116$

$P(x_1|C_2) = \frac{1}{2\pi \times \sqrt{0,215}} e^{-\frac{1}{2}\left(([1\ 0]-[0,755\ 0,866]\begin{bmatrix}0,654 & -1,044\\ -1,044 & 1,368\end{bmatrix}^{-1}\left(\begin{bmatrix}1\\0\end{bmatrix}-\begin{bmatrix}0,754\\0,866\end{bmatrix}\right)\right.}$

$= 0,149$

$\gamma_{11} = \frac{0,116 \times 0,393}{0,116 \times 0,393 + 0,149 \times 0,607} = 0,335$

$\gamma_{21} = \frac{0,149 \times 0,607}{0,149 \times 0,607 + 0,116 \times 0,393} = 0,665$

$P(x_2|C_1) = \frac{1}{2\pi \times \sqrt{0,125}} e^{-\frac{1}{2}\left(([0,2]-[2,22\ \ -0,45])\begin{bmatrix}1,150 & -0,861\\ -0,861 & 0,728\end{bmatrix}^{-1}\left(\begin{bmatrix}0\\2\end{bmatrix}-\begin{bmatrix}2,22\\-0,45\end{bmatrix}\right)\right)}$

$= 0,001$

$P(x_2|C_2) = \frac{1}{2\pi \times \sqrt{0,215}} e^{-\frac{1}{2}\left(([0,2]-[0,755\ 0,866])\begin{bmatrix}0,654 & -1,044\\ -1,044 & 1,368\end{bmatrix}^{-1}\left(\begin{bmatrix}0\\2\end{bmatrix}-\begin{bmatrix}0,755\\0,866\end{bmatrix}\right)\right.}$

$= 0,207$

$\gamma_{12} = \frac{0,001 \times 0,393}{0,001 \times 0,393 + 0,207 \times 0,607} = 0,003$

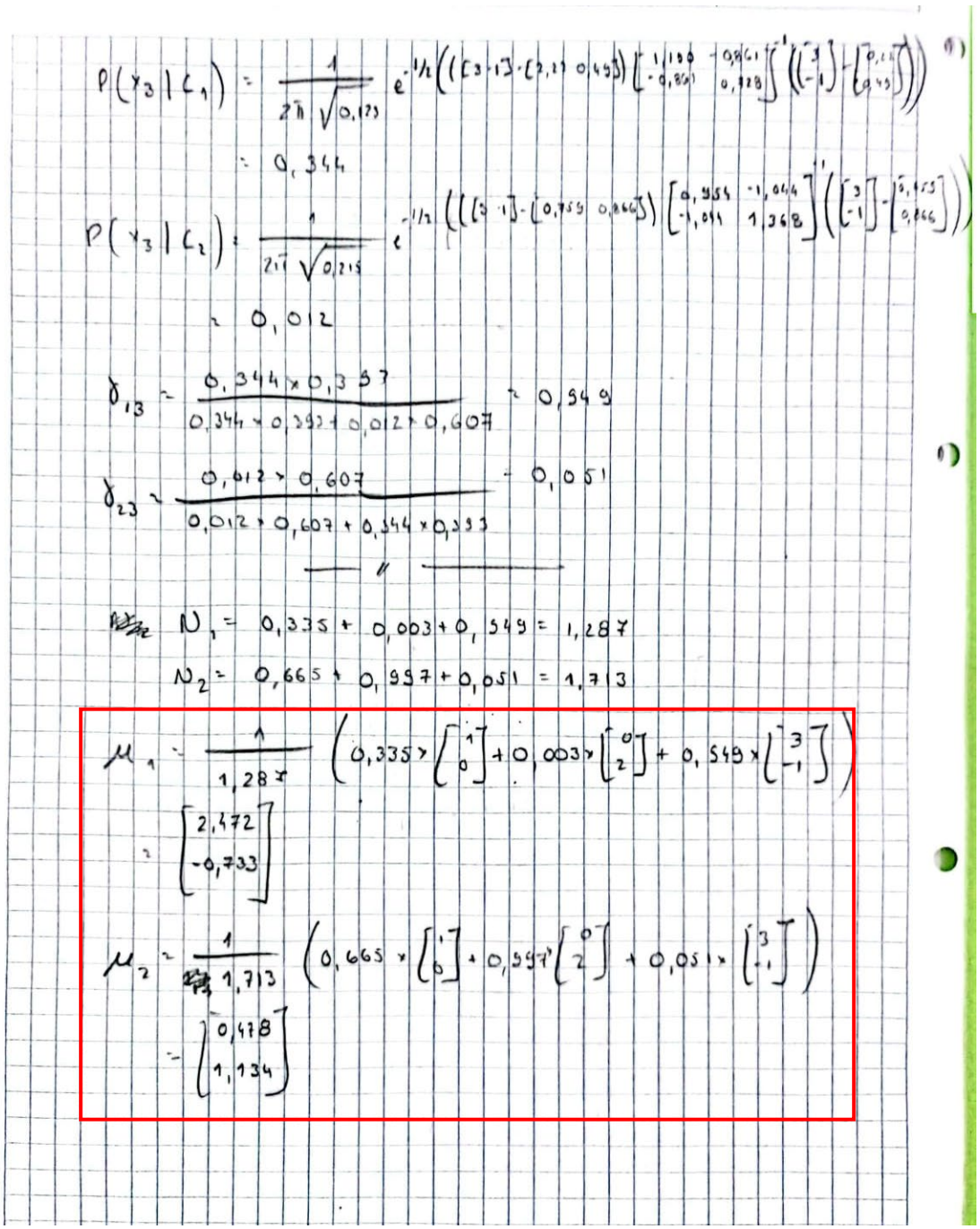
$\gamma_{22} = \frac{0,207 \times 0,607}{0,207 \times 0,607 + 0,001 \times 0,393} = 0,997$

TÉCNICO LISBOA

Aprendizagem 2024/25

Homework IV – Group 066

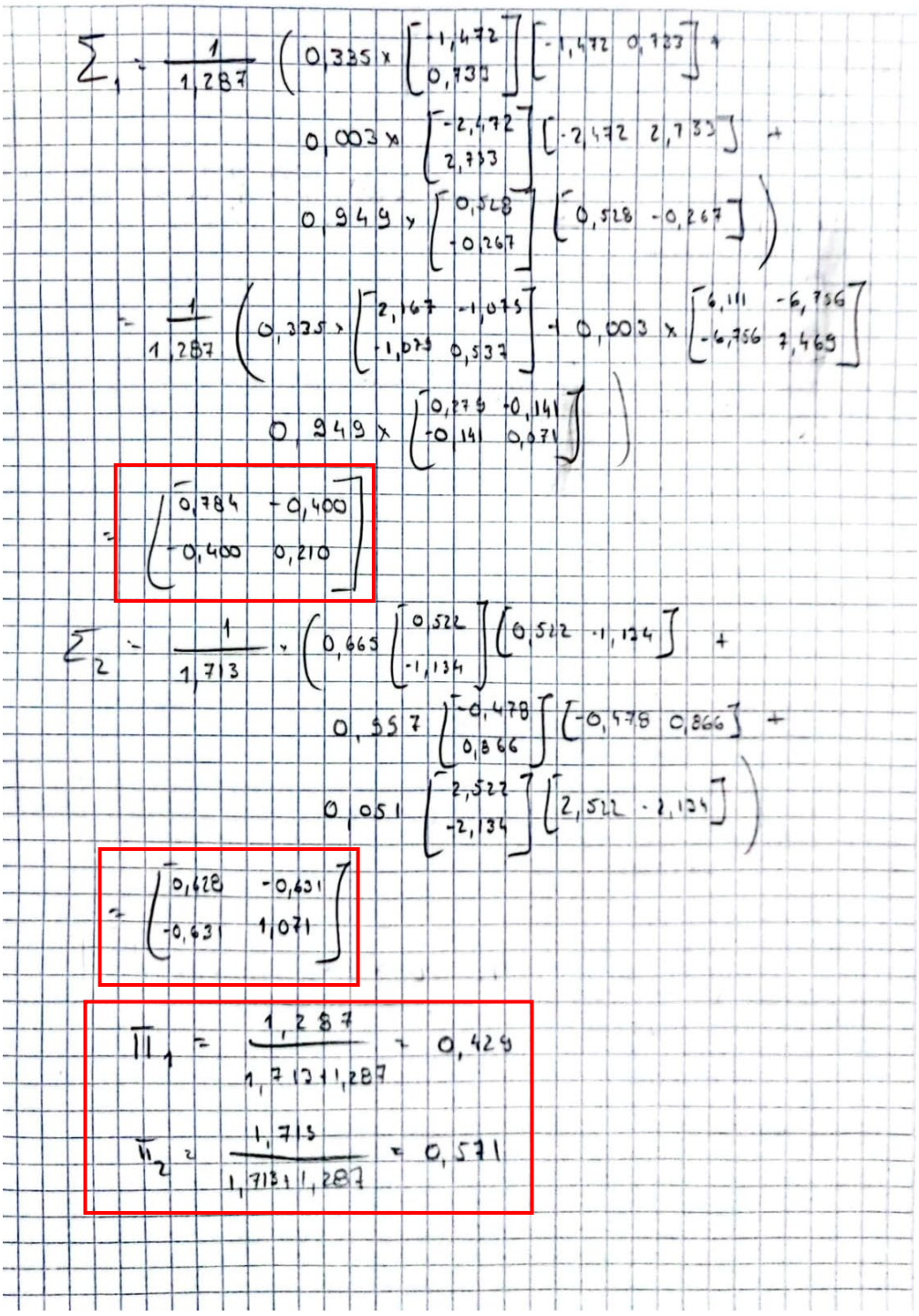
(ist106794, ist1107301)



Aprendizagem 2024/25

Homework IV – Group 066

(ist106794, ist1107301)



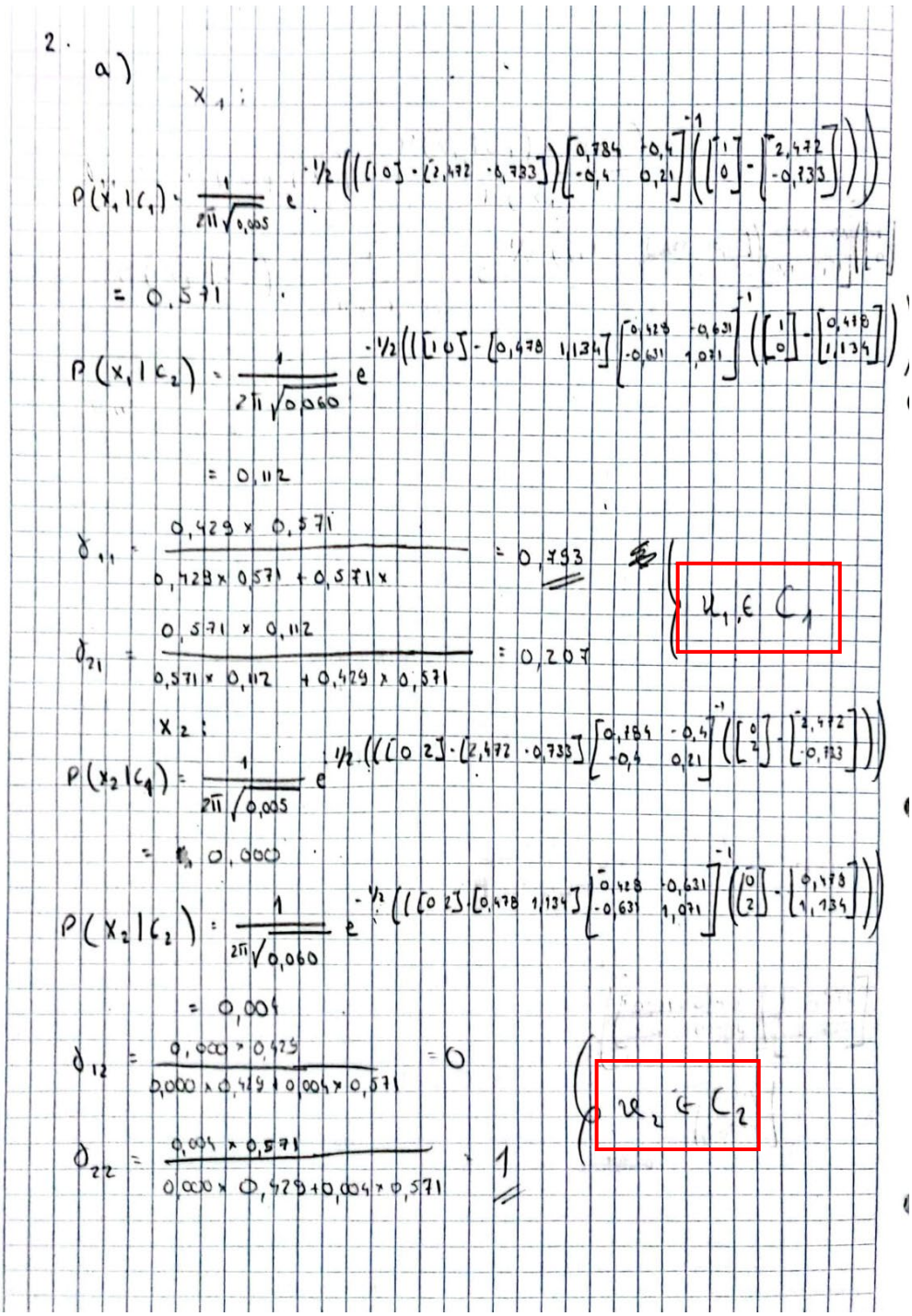
Aprendizagem 2024/25

Homework IV – Group 066

(ist106794, ist1107301)

2)

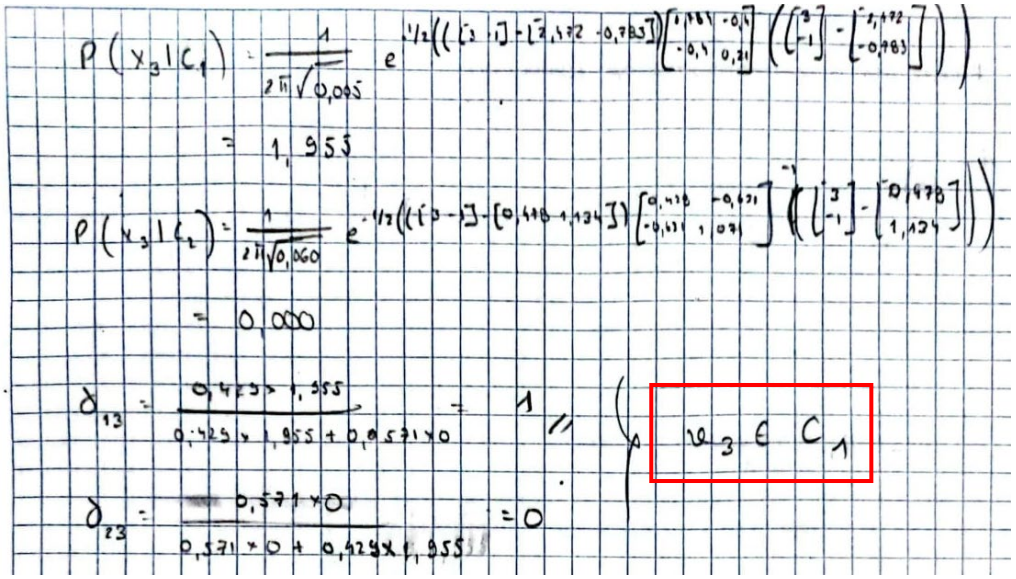
a)



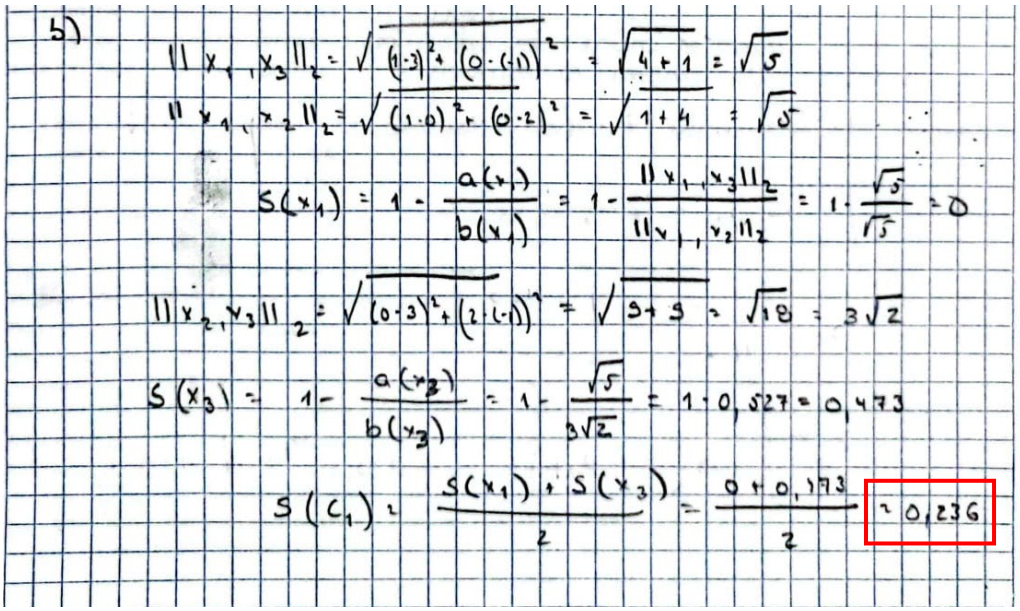
Aprendizagem 2024/25

Homework IV – Group 066

(ist106794, ist1107301)



b)



Homework IV – Group 066

(ist106794, ist1107301)

II. Programming and critical analysis

Here we have an overview of all the imports used and how we imported the dataset, which are common to all questions. Therefore, we will be omitting this code from the beginning of each task to avoid redundancy

1)

```
import pandas as pd, matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA

# Load the data
df = pd.read_csv('accounts.csv')
X = df.drop('deposit', axis=1)
X = X.iloc[:, :8].drop_duplicates().dropna()
X = pd.get_dummies(X, drop_first=True)
Y = pd.get_dummies(df['deposit'], drop_first=True)
```

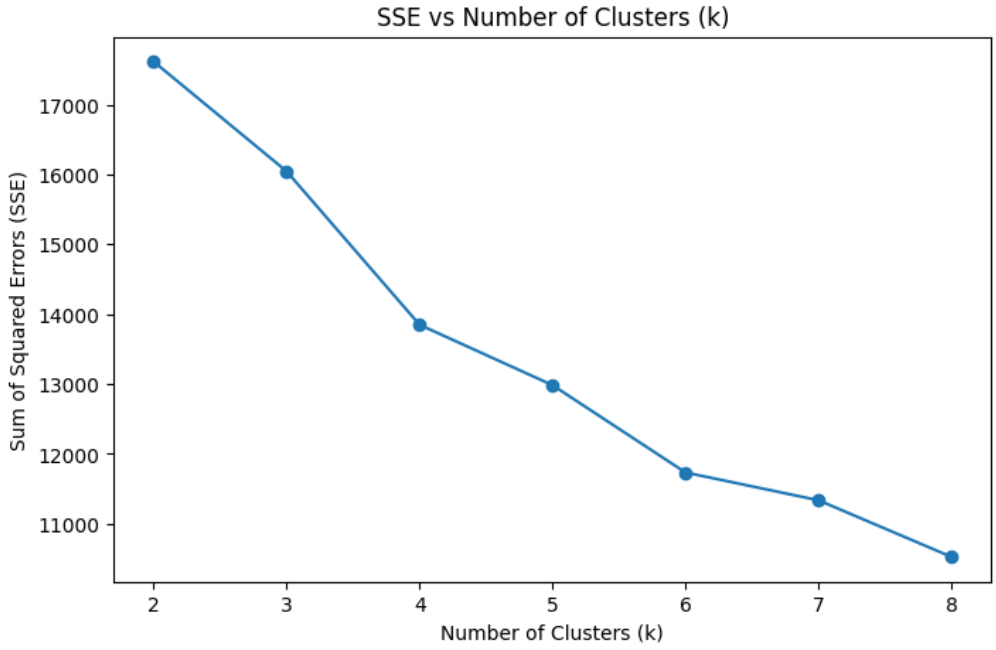
a) Normalizing the data and applying k-means for each k value

```
# Normalize the data using MinMaxScaler
scaler = MinMaxScaler()
X_normalized = scaler.fit_transform(X)

sse = []
k_values = range(2, 9)

# Apply k-means clustering for each k value
for k in k_values:
    kmeans = KMeans(n_clusters=k, max_iter=500, random_state=42)
    kmeans.fit_predict(X_normalized)
    sse.append(kmeans.inertia_)
```

Plotting the result

```
# Plot the SSE for each k
plt.figure(figsize=(8, 5))
plt.plot(k_values, sse, marker='o')
plt.title('SSE vs Number of Clusters (k)')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Sum of Squared Errors (SSE)')
plt.show()
```

Output:



b) According to the plot, and using the knee/elbow finding method we come to the conclusion that the ideal number of clusters is 4. This method consists of finding the point where the decrease in SSE becomes less accentuated with the increase of clusters.

For all values before this point, the model might suffer from underfitting and for values above, the increase in the number of clusters and complexity does not provide substantial improvements to the model. Not only does it not justify, it also contributes to overfitting the data.

c) K-modes is an adaptation of k-means used to better handle categorical features. It uses the mode (most frequent value) instead of means to represent the centroid of the cluster and uses distance based in dissimilarity (like Hamming Distance) to group the data.

Given that our dataset's features are predominantly categorical (10 categorical out of 17), in theory, k-modes would be a better clustering approach, considering the explanation above.

2)

a) Normalizing the data with StandardScaler, applying the PCA and getting the explained variance ratio for each component

```
scaler = StandardScaler()
X_normalized = scaler.fit_transform(X)

# Apply PCA
pca = PCA(n_components=2)
pca.fit(X_normalized)

variance_explained = pca.explained_variance_ratio_

print("The first component explains", round(variance_explained[0], 5), "variability in the data set.")
print("The second component explains", round(variance_explained[1], 5), "variability in the data set.")
print("The top 2 components explain", round(variance_explained[1] + variance_explained[0], 5), "variability in the data set.")
```

The first component explains 0.11679 variability in the data set.

The second component explains 0.11076 variability in the data set.

The top 2 components explain 0.22755 variability in the dataset.

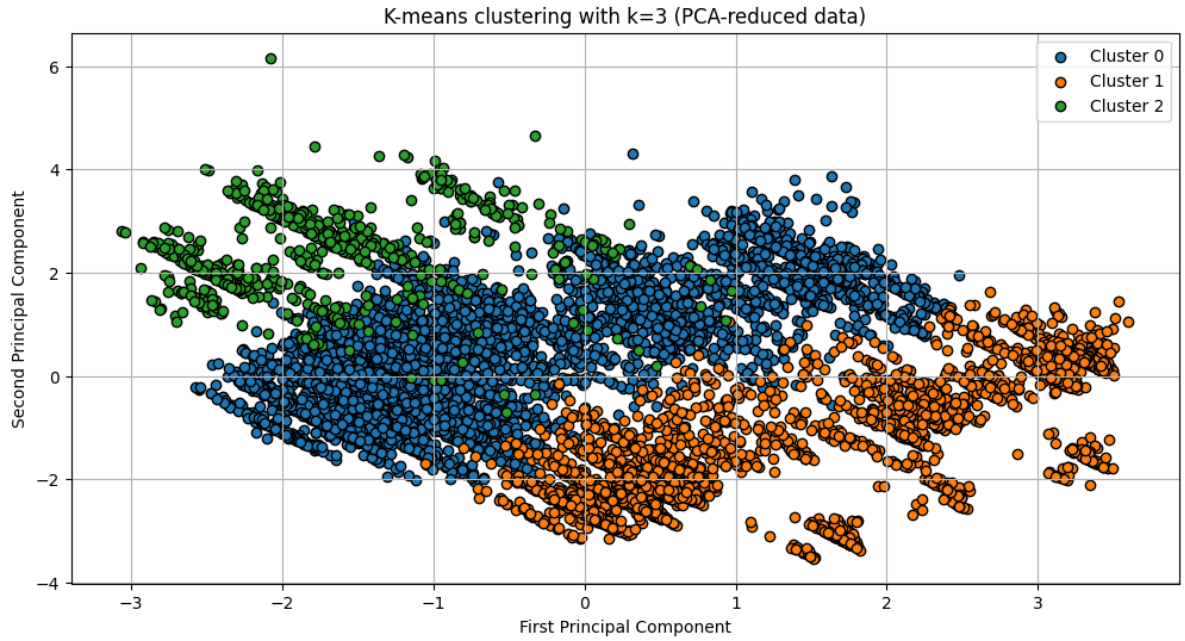
b) Applying k-means clustering and PCA

```
kmeans = KMeans(n_clusters=3, random_state=42)
labels = kmeans.fit_predict(X_normalized)

X_projected = pca.fit_transform(X_normalized)
```

Plotting the results:

```
plt.figure(figsize=(12, 6))
for cluster in range(3):
    plt.scatter(X_projected[labels == cluster, 0], X_projected[labels == cluster, 1], label=f'Cluster {cluster}', edgecolors='k')

plt.title('K-means clustering with k=3 (PCA-reduced data)')
plt.xlabel('First Principal Component')
plt.ylabel('Second Principal Component')
plt.grid()
plt.legend()
plt.show()
```

Output:



We can clearly separate all the three clusters. There is still a bit of overlap between some of them but most of the points are relatively close to their supposed cluster

The overlap can happen because we are losing a lot of information on the data variance by only considering the top two components to draw the plot. It can suggest that using only these two components in a 2D space isn't enough to represent some of the variance between the different observations of the dataset, that might become more apparent in higher dimensional spaces.

c) Plotting both graphs

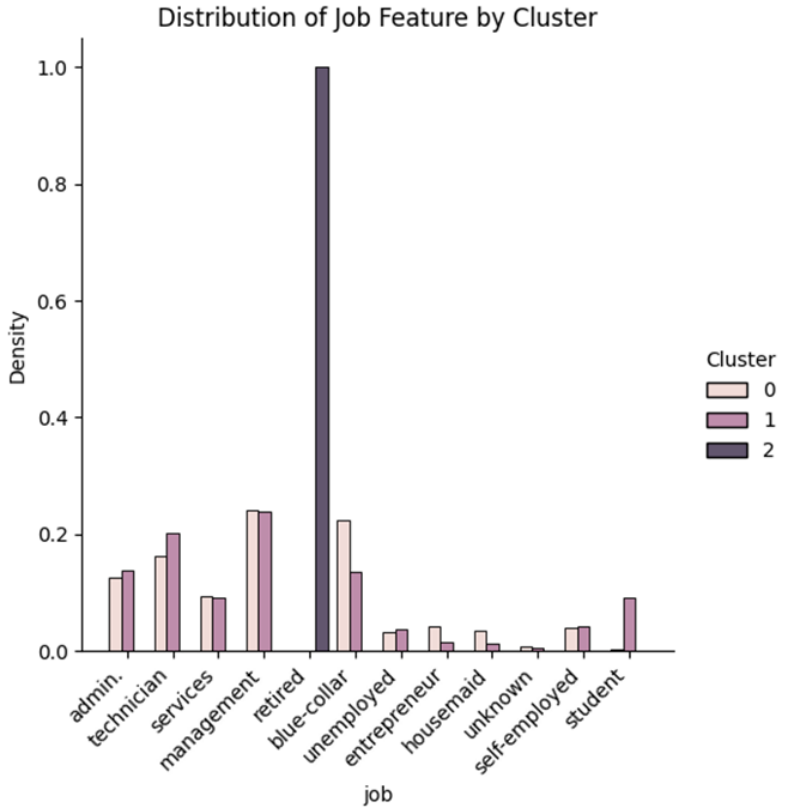
```
#Getting the dataframe with the job and education columns (get_dummies removes such column's names)
df_labeled = pd.DataFrame(df[['job', 'education']]).loc[X.index]
df_labeled['Cluster'] = labels

# Plot the 'job' feature distribution across clusters
sns.displot(df_labeled, x="job", hue="Cluster", multiple="dodge", stat='density', shrink=0.8, common_norm=False)
plt.title('Distribution of Job Feature by Cluster')
plt.xticks(rotation=45, ha = 'right')
plt.show()

# Plot the 'education' feature distribution across clusters
sns.displot(df_labeled, x="education", hue="Cluster", multiple="dodge", stat='density', shrink=0.8, common_norm=False)
plt.title('Distribution of Education Feature by Cluster')
plt.xticks(rotation=45)
plt.show()
```

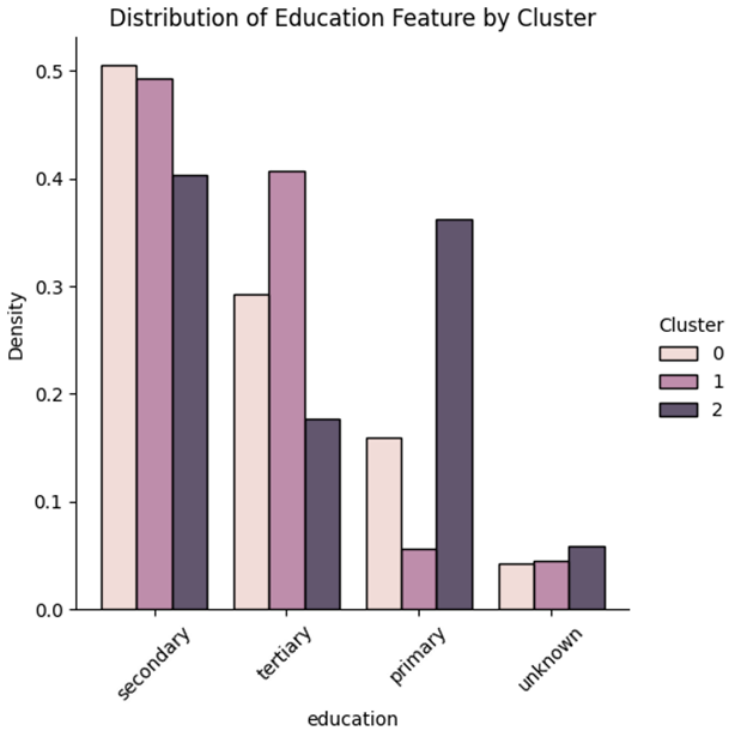
Aprendizagem 2024/25

Homework IV – Group 066

(ist106794, ist1107301)

Output:





The main differences in the obtained plots can be described as follows:

- Jobs
  - Cluster 0: Jobs like “services”, “management”, “blue-collar”, “entrepreneur” and “housemaid” appear mostly in this cluster. These jobs usually require less technical knowledge and might translate to a lesser income than some others in this list.
  - Cluster 1: Jobs like “technician”, “admin” and “student” appear mostly in this cluster. This can indicate that a certain level of education and technical knowledge is common to the people that are part of this cluster’s population.
  - Cluster 2: Fully represented by people in the “retired” category. In terms of income, it is represented by a demographic of people with fixed income.

- Education
  - Cluster 0: People with “secondary” level of education appear in a higher portion in this cluster. It represents a good correlation with the results in the first graph, where this cluster is associated with less technical jobs and lower income levels.
  - Cluster 1: People with “tertiary” level of education appear mostly in this cluster. There’s also a big portion of people with “secondary” level. Once again, this graph further explains the association between this cluster and more technical jobs with higher income.
  - Cluster 2: Highly characterized by people with “primary” level of education and also some of “secondary” level. This distribution fits perfectly with the idea of older and “retired” individuals, who mostly entered the job market with smaller education levels when compared to today’s generation.

END