DISS - Machine Learning (0 points)
December 2024

All documents allowed. Read all directions carefully and write well-argued answers.
Try to be concise but precise.
You must write your answers in english, especially for the part of Andrea Failla.

1 Section by Rémy Cazabet (10)

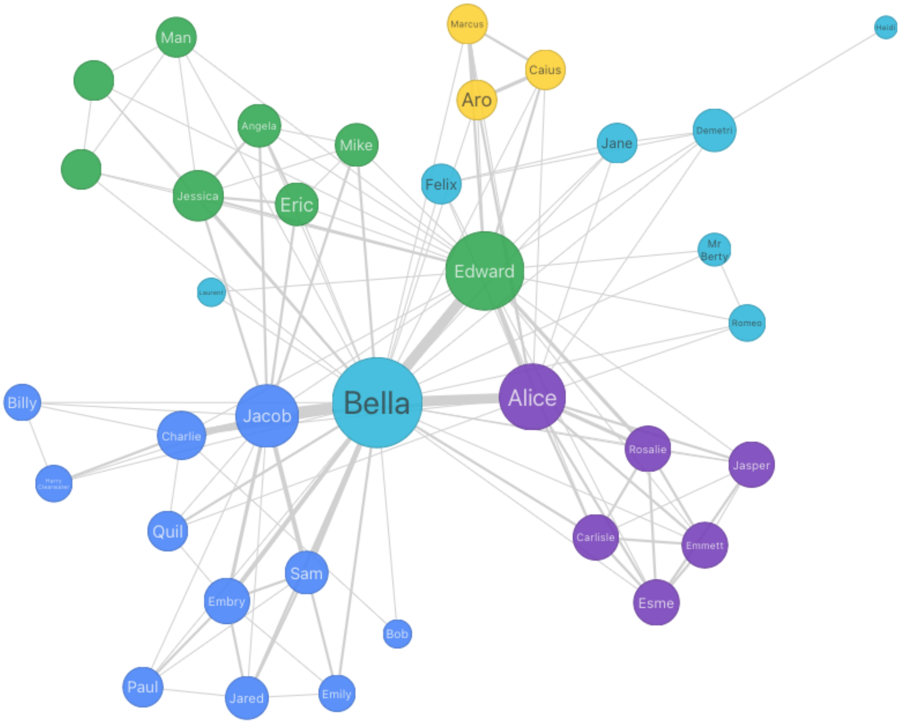

Figure 1: Movie's network

[2] Look at the network in Fig.1. Would you say that this network is a small world network ? Explain. You do not need to give quantitative values.

[2] What do you think are the most important nodes on this network, using the concept of centrality?

[4] You are working for a public transport company seeking to analyze user travel habits to improve its services. You have access to a dataset containing information for each trip made over a given period, including: user ID, departure time (string, "DD/MM/YYYY"), trip duration, origin and destination (name of train station), type of subscription (monthly, weekly, single ticket), and the total number of trips made by the user in a month.

One of the objectives is to identify user profiles, i.e., groups with similar travel habits.

1. By examining the available variables, do you think it is necessary to preprocess them before running a clustering method on them? If yes, which ones, and why?

2. How would you proceed to perform the grouping ?

[2] You are working for a sports equipment store that wants to predict the sales price of running shoes based on their features. The dataset contains the following variables for each pair of shoes:

- **Brand** (e.g., Nike, Adidas, Puma)
- **Weight of the shoe** (in grams)
- **Cushioning level** (Numerical value)
- **Material type** (synthetic, leather, mesh)
- **Price** (in euros)

The store's objectives are:

1. Develop a model that can generalize well to predict prices of new shoes, especially for cushioning levels or weights outside the range of the training data.
2. Accurately predict prices for shoes within the range of the training data, even if this involves capturing complex interactions between features (e.g., the effect of cushioning level depending on the material type).

What method should they use for their predictions, and why?