**Project on the application of statistical methods for AI**

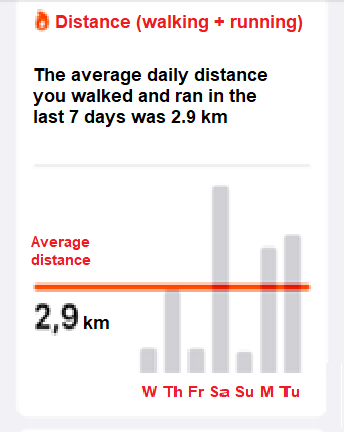
This project will be individual and collective as well, concerning the entire student group. It aims to increase diversity and a sense of responsibility for the quality of their data. Its essence consists in the semester-long collection of measurable (numerical) and unmeasurable data, often only binary 0 and 1. It is important that the data is collected systematically, e.g. in Excel in the form of a table.

Both projects do not have clearly defined boundaries, because only during the classes students learn various statistical methods that allow for a deeper penetration of the collected data.

In the individual project, this opens up space for launching your own creativity supported by the suggestions of the GPT chat (I accept it). The more threads are added and developed, the better.

**Individual project**

It consists in recording the daily length of the walking route while walking. You can use your own smartphone equipped with tools for such measurements for this purpose. For example, the Health app on the iPhone will show the number of steps and/or kilometers (you may need to activate it to count the mileage).

If daily registration, i.e. precise recording of the distance is not done, for example, due to forgetfulness, there is a chance to read these readings from the weekly overview, but without indicating specific values. Then everyone should approximate them themselves, relating the height of the bar to the average value shown in the graphic. For example, the highest bar in the graphic below refers to Saturday and indicates a distance of about 6.7 km. I used a millimeter ruler and the proportion:

2,9 km ~ 20 mm

*x* km ~ 49 mm

Each student fills in an Excel table with the following characteristics (features)

|gender F/M| M | Tu | W | Th | Fri | Sa | Su | calories burned |

entering the necessary data daily. The research goal is identical to that in the collective project presented below, and the number of methods used will be a measure of knowledge acquisition in this area of ​​teaching.

**Collective project**

This involves collecting data on store purchases and recording the amounts spent (without revealing the assortment of the purchase). This requires keeping a receipt and recording the total of daily expenses, so that they can then be added up for a given day. According to the calendar, there are 62 days left until January 10 (the end of the project), which is enough to collect reliable and representative data. It will also be necessary to record whether the purchase was made on a working day (5 days) or on weekends (2 days). The data must also include information on the gender of the person collecting the data and the district of residence.

**Research objective**

The research objective is to build statistical models and apply methods such as:

* **linear regression** (one- and multi-dimensional) to find any relationships between the amount of money and the gender of the person, or the fact that the purchase was made during the working week or at the weekend,
* **logistic regression** to assess whether the gender of the person can be deduced from the amount of expenditure,
* **k-means** used for so-called clustering, i.e. discovering hidden structures in data and can be used in customer analysis, image segmentation, analysis of user behaviour in many other areas where grouping objects with similar features is important. This is an action involving grouping by yet unknown features.
* **k-NN** (**k - Nearest Neighbours**) used in turn to classify data points and determine their affiliation to a category. There are no a priori given features here, and at most, we can identify them and only then name them. This is a method that follows logistic regression, but there are known categories (features), such as the gender of the person.
* **testing statistical hypotheses** assessing the significance of differences between categories (e.g. whether women spend more money than men on average),
* **naive Bayesian** and/or **random forest** - also used to classify data according to specific features.

When all the data is collected (right after the winter break), the expenses for the entire observation period will be averaged during the exercises and returned to the students. Then you can also start using the above methods as work on a joint project.