Introduction:

In this project, you have the opportunity to employ your data mining knowledge in a real-world problem. To this end, only a data set is provided to you. Your job is to analyze this dataset to find interesting problems and finally to solve them correctly. In other words, you must work on the dataset from the initial steps (that can be visualization) until the last steps (that can be evaluation of the models you have trained). Accordingly, this is a comprehensive project that accounts for 75% of your grade.

You should do the project in groups of two students, but if you'd like to do it individually, it's also possible.

Project Tasks:

For this project, this is up to you to explore the data set and come up with interesting Data Mining problems. Based on those interesting problems, you should define concrete, well-formulated tasks. For this project, you must define at least two supervised machine learning tasks and one unsupervised machine learning task. You need to describe the defined tasks clearly in the report, because this is the crucial part that's a foundation for the whole project.

It is also worth noting that you must start with putting some effort on data visualization and exploratory data analysis before you go to algorithms and AI techniques. It helps a lot in getting acquainted with the dataset, and will be instrumental in defining good tasks.

At least one of your tasks must include identifying and addressing an interesting and challenging problem, where simply using out-of-the-box algorithms would not work. For example, it could be class imbalance, feature extraction/selection, label noise, etc. You should then design a non-trivial solution, argue for it, and finally evaluate it.

Once you have defined your tasks, you must solve them correctly. When solving data mining problems, there are several perspectives to consider, including:

- What are the challenges that you are faced with (whether related to instances, features, labels, etc.)?
- What are the solutions you use to address the challenges and to improve the results (can be data balancing, feature selection, feature extraction, etc)?
- What are the selected methods/algorithms for supervised and unsupervised ML techniques? What is your motivation for choosing them? What are the weaknesses and strengths of them, in the context of your tasks? What are the considerations that must be taken into account for them (you may apply multiple different algorithms for your tasks)?
- How do you evaluate the quality of your result (it is important that you do a rigorous evaluation with a statistical significance test, etc)?

Data description:

In this project, you will work on data about some vehicles. These vehicles are hybrid but they can switch between and operate in the electric, diesel, and hybrid modes.

There are two versions of data that you can use whichever version you prefer.

Version 1: Measurements in SENDTIME+Vehicles' information
You can use Measurments_In_SENDTIME.csv and VehiclesInformation.csv files together. In both files, vehicles are identified by "Vehicle_ID".

Measurments_In_SENDTIME.csv contains measurements of the vehicles that are calculated in SENDTIME. For example, measurements that show the state of health of the battery of the vehicle, the amount of time that the vehicle operates in each mode, the distances traveled by the vehicle and etc. Although for each vehicle there may exist a number of measurements, the measurements are not provided regularly, so the time interval between consecutive samples is not fixed.

Vehiclesinformation.csv contains information about vehicles that are specified by "Vehicle_ID". These are about the specification of the vehicles such as the country or geographical area in which the vehicle is used, information about the battery including the generation of the battery as well as the date that the battery is replaced, the mounted battery generation, battery supplier and etc.

Version 2: Project_Dataset
Measurements in SENDTIME.csv and Vehicles' information.csv files are merged together using the "Vehicle_ID" key. So, the two provided versions contain same information.

The description of the features in the datasets are provided in the **Dataset Description** file.

It is noteworthy that this dataset has been anonymized. So, for example, you can not see the names of countries, instead, the name of each country is encrypted with a number.

Report:
Report should be structured like a research paper: presenting your goals, explaining your methods, reporting on your experiments, and finally summarizing your findings and drawing conclusions.

It does not have to be very long, however, it needs to be sufficiently detailed to make it clear what you have tried (and to include justification and explanations for your thought process). The project is the most important part of this course, so please make it clear how you spent the 200 hours of studying that corresponds to its 7.5 credits.

We are expecting around 5 pages of text (so the full report will probably be 3-5 times as long since it is expected to include a lot of result plots and tables).

The deadline for submitting the report is Friday, 23 October 2020 (however, if you want to have the individual meetings early in week 44, and make the public presentation, you should submit your report at least a couple of days earlier)

Peer review:
The report from each group would be read and commented on by another group. You will receive your colleagues' comments before your presentation.

Presentation:
We will have individual meetings with each group to discuss the project report throughout week 44, about 20 minutes long. You should prepare a 2-3 minutes summary/presentation of your work (no slides!) – it's very short, so focus on one or two most interesting findings. We will then have a discussion, to evaluate your level of understanding and, potentially, individual contributions to the final report.
If you want grade 5, you will also need to make a public presentation on Friday, 30 October. It'll be 15 minutes to show your colleagues (and course teachers) what you have done in the project. Explain the problem that you have decided to tackle, the approach that you are proudest of, and why you have decided to proceed in this way.

Evaluation Criteria:
You will be graded based on three main criteria:
1. how comprehensive your work is, which means basically the more methods you try and the better you develop them, the better grade you get;
2. how deep your understanding of Data Mining approaches is -- it is important to show whether you can justify your choice of techniques for the problems that you have addressed;
3. and your ability to design experiments and interpret their results, which is revealed by whether your conclusion is supported and justified by your results.