INF264
Project 2:
Predicting traffic

Deadline: October 9th, 23.59
Deliver here:
`https://mitt.uib.no/courses/24958/assignments/32001`

Projects are a compulsory part of the course. This project contributes a total of 15 points to the final grade. You need to upload your answer to MittUiB before 23.59 on the 9th of October.

The project can be done either alone or in pairs. If you do the work with a pair, add a paragraph to your report explaining the division of labor (Note that both students will get the same grade regardless of the division of labor).

Grading will be based on the following qualities:

- Correctness (your answers/code are correct and clear)
- Clarity of code (documentation, naming of variables, logical formatting)
- Reporting (thoroughness and clarity of the report)

Especially, weight is put to correct use of model selection and evaluation procedures.

Deliverables. You should deliver exactly two files:

1. a PDF report containing an explanation of your approach and design choices to help us understand how your particular implementation works. You can include snippets of code in the PDF to elaborate any point you are trying make.

2. a zip file of your code. We may want to run your code if we feel necessary to confirm that it works the way it should. Please include a README.txt file in your zip file that explains how we should run your code. In case you have multiple files in your code directory, you must mention in the README.txt file which file is the main file that we need to run to execute your entire algorithm.

Programming languages. You are allowed to submit your implementation in the following languages: Python, Matlab, C/C++, Java. (This list is based on the skills of the graders and no new languages will be added later.)

Code of conduct. You are allowed to use existing libraries. However, **submitting code that is written by someone else is considerer cheating**. If you are unsure whether something is allowed or not, ask the teaching assistants.

Late submission policy: All late submissions will get a deduction of 2 points. In addition, there is a 2-point deduction for every starting 12-hour period. That is, a project submitted at 00.01 on October 10th will get a 4-point deduction and a project submitted at 12.01 on the same day will get a 6-point deduction (and so on). All projects submitted on October 12th or later are automatically failed. (Executive summary: Submit your project on time.) There will be no possibility to resubmit failed projects so start working early.

1 Task: Predicting traffic

To goal of this task is to construct a model that predicts traffic (number of cars) crossing Gamle Nygårdsbro at a given time.

1.1 Data

Statens vegvesen has a traffic monitoring station at Gamle Nygårdsbro right next to the Department of Informatics. In this project, we use hourly traffic data from December 2015 to December 2019. We are going to predict three different quantities: The total number of cars crossing Gamle Nygårdsbro (`Volum totalt`), the number of cars towards city center (`Volum til SNTR`), and the number of cars towards Danmarksplass (`Volum til DNP`). The features are year (`År`), month (`Måned`), day (`Dag`) and time (`Fra_time`).

The data file `data.csv` can be downloaded from MittUiB.

1.2 Preprocessing

The first step in the modelling process is to preprocess your data so that you can input it to a learning algorithm. To this end, you need to extract some features from the date.

Remember feature engineering. Use your domain knowledge about the traffic patterns. What kind of variations there are?

Think whether a feature should have a continuous or categorical value. It may make sense to represent categorical features with more than two possible values using one-hot encoding.

Hint: Visualize data to see what kind of patterns there are.

1.3 Modelling and evaluation

Learn at least three models that predict the total number of cars crossing Gamle Nygårdsbro during one hour given the date and time; remember to try several different hyperparameters. Select the best model. Estimate the expected performance on unseen data of the chosen.

Repeat the previous procedure but this time predict the number of cars heading towards Sentrum. Next, predict the number of cars heading towards Danmarksplass.

How do the patterns towards Sentrum and Danmarksplass differ? Are the models same?

When is your model good/bad? What are limitations of your model?

Hint: Visualize your predictions. Perform sanity checks (Does your model make predictions that are obviously wrong?).

Suppose you deployed the system in the beginning of year 2020. Would your model have performed as expected? Can you explain why/why not? What did we learn from this? (You can use the data in file `data_2020.csv`)

1.4 Bonus task

This task is not compulsory.

Find additional data and use it make a better model.