IML 2023: Predicting Saturation Vapour Pressure From Molecular Properties

Gauri Pradhan & Rafael Savvides & Kai Puolamäki

2023

DATA11002 Introduction to Machine Learning (Autumn 2023)

In this term project, you will train a regression model on a data set of atmospheric measurements. To complete the project, you should deliver:

- **Sun 10 December:** Predictions for the test set are to be submitted on Kaggle with the link sent to the students who have passed the prerequisite examination, and a preliminary version of your project report as a single PDF file in Moodle.
- **Fri 15 December:** Term project presentation for some of you.
- **Sat 23 December:** The final report as a single PDF file in Moodle.

About the data

The term project is based on the **GeckoQ** dataset with atomic structures of 31,637 atmospherically relevant molecules resulting from the oxidation of α-pinene, toluene and decane. The GeckoQ dataset is built to complement data-driven research in atmospheric science. It provides molecular data relevant to aerosol particle growth and new particle formation. A key molecular property related to aerosol particle growth is the saturation vapour pressure (pSat), a measure of a molecule's ability to condense to the liquid phase. Molecules with low pSat, low-volatile organic compounds (LVOC) are particularly interesting for NPF research. All the data in GeckoQ pertains to LVOCs. (For more, read: *Besel et al.* https://doi.org/10.1038/s41597-023-02366-x).

For each molecule, GeckoQ features important thermodynamic properties: *saturation vapour pressure* [Pa] (pSat), *the chemical potential* [kJ/mol], the *free energy of a molecule in mixture* [kJ/mol], and the *heat of vaporisation* [kJ/mol]. Out of these, the **saturation vapour pressure** will be the focus of your term project. There are two types of features that you will have the choice of using for your project: the interpretable features (described in detail below) and the topographical footprints (TopFP) of the molecules. Previous works record using the TopFP descriptor as inputs to a machine learning model to learn pSat as a function of atomic structure for a different dataset. (*Wang et al.* https://doi.org/10.1073/pnas.1707564114)

Following are the columns which make up the training/test data set. Barring the `Id` and the `pSat_Pa` columns, all others form the interpretable features of the molecules:

- `Id` - A unique molecule index used in naming files.
- `MW` - The molecular weight of the molecule (g/mol).
- `pSat_Pa`- The saturation vapour pressure of the molecule calculated by COSMOtherm (Pa).
- `NumOfAtoms` - The number of atoms in the molecule.
- `NumOfC` - The number of carbon atoms in the molecule.
- `NumOfO`- The number of oxygen atoms in the molecule.

- NumOfN- The number of nitrogen atoms in the molecule.
- NumHBondDonors - "The number of hydrogen bond donors in the molecule, i.e. hydrogens bound to oxygen."
- parentspecies- Either "decane", "toluene", "apin" for alpha-pinene or a combination of these connected by an underscore to indicate ambiguous descent. In 243 cases, the parent species is "None" because it was not possible to retrieve it.
- NumOfConf- The number of stable conformers found and successfully calculated by COSMOconf.
- NumOfConfUsed- The number of conformers used to calculate the thermodynamic properties.
- C = C (non-aromatic)- The number of non-aromatic C=C bounds found in the molecule.
- C = C-C = O in non-aromatic ring- The number of "C=C-C=O" structures found in non-aromatic rings in the molecule.
- hydroxyl (alkyl) - The number of the alkylic hydroxyl groups found in the molecule.
- aldehyde- The number of aldehyde groups in the molecule.
- ketone - The number of ketone groups in the molecule.
- carboxylic acid - The number of carboxylic acid groups in the molecule.
- ester - The number of ester groups in the molecule.
- ether (alicyclic)- The number of alicyclic ester groups in the molecule.
- nitrate - The number of alicyclic nitrate groups in the molecule.
- nitro - The number of nitro ester groups in the molecule.
- aromatic hydroxyl - The number of alicyclic aromatic hydroxyl groups in the molecule.
- carbonylperoxynitrate - The number of carbonylperoxynitrate groups in the molecule.
- peroxide - The number of peroxide groups in the molecule.
- hydroperoxide- The number of hydroperoxide groups in the molecule.
- carbonylperoxyacid- The number of carbonylperoxyacid groups found in the molecule
- nitroester- The number of nitroester groups found in the molecule

Your task

You should work in groups of 1-3 students.

Since saturation vapour pressure is a continuous variable, your task is to build a regression-based machine-learning model that uses the aforementioned interpretable features or topographical fingerprints of the molecules.

NOTE: This is a non-trivial regression task. It is possible to do it in many ways. The most straightforward regression model that you could build is a linear regressor, but that is inefficient for the task since it assumes a linear relationship between input features and the predicted values. Therefore, you should undertake thorough data exploration, pre-processing, feature selection, model selection, R2 score estimation, etc., appropriately since you will report and analyse your choices and results in the project deliverable.

The project's purpose is not to (even try to!) replicate any methods in the literature, make a super-complex best-performing classifier that beats everything else or attempt to use other data sources, etc., to obtain the best possible R2 score . Do not use any method that you do not understand yourself! Accuracy of the predictions on the test data is not a grading criterion by itself, even though a terrible accuracy may indicate something else fishy in your approach (which could affect grading).

The Online Challenge (DL 10 December)

We are organising a non-serious competition (or "challenge") to make the project more interesting. We will be using the R2 score (https://en.wikipedia.org/wiki/Coefficient_of_determination) as a metric to evaluate your submission, all computed by comparing your predictions on the test data to the correct labels (which we have, but you don't). The R2 score is essentially a measure of correlation between the model's predictions and the ground truth. So, the higher the R2 score, the better, but it can also be negative if the model chosen to fit the data is a poor choice for the problem.

Since we are using Kaggle to collect your submissions, we will use a subset of test samples for the private leaderboard. For those unfamiliar with Kaggle, it's a submission's score on the private rows that will be used to determine the final standings. This "private leaderboard" is only viewable by the competition host until the competition deadline, after which we publish it for participants.

At the latest, by 10 December, you should submit the following:

- your predictions to Kaggle, and
- a preliminary version of your report to Moodle.

After this deadline, we will publish the private leaderboard scores of every team on Kaggle (and the correct labels for the test set will be made available on Moodle).

The preliminary report should describe the work done so far. This report does not need to be polished or complete, but it should already contain the basic ideas used in the solution. The teams are allowed to modify their approach and report before they submit their final report by 23 December. However, please do not simply copy the method used by the teams with good performance in the competition!

The Final Report (DL 23 December)

You should submit the final report as a PDF file via Moodle by 23 December.

The final report should contain, among other things, the following:

- The names of the group members.
- The name of the team you used to submit the predictions on Kaggle.
- The stages of your data analysis, including how you looked at the data to understand it (visualisations, unsupervised learning methods, etc.).
- Description of considered machine learning approaches and pros and cons of the chosen approach for this application.
- Steps you took to select good features and model parameters.
- Summary of your results, insights learned, and how the regression model performed.
- As a final section, please include a self-grading report (at most 1 page) that suggests a grade for yourself (integer 0-5) by using the attached grading instructions (see below).

It is enough to use one of the basic algorithms, do the feature and model selection parts as instructed (you should probably use cross-validation!), and prepare a well-written report to pass the project.

Practical instructions for writing the report:

- Your report should read like a self-contained blog post or scientific article that is understandable and without any task description. You should explain what you have done and why you have done it so that a person familiar with machine learning can understand what you have done and could, in principle, reproduce what you have done based on your report alone. Put some emphasis on presentation and readability (one of the grading criteria): imagine that the report's reader would be your future boss, who appreciates a clear and concise presentation.

- You are not required to hand in any program code. Your report, thus, should look different from a code listing! Your report may contain code snippets if you explain what the reader is supposed to conclude from your code. We may look at them, but we won't go fishing for results and missing details from your code. In other words, all relevant parts of your report should be understandable without going through any code. If you need to include more significant chunks of code, please put them in an appendix so we can easily skip them when grading your report.

- Your report may include tables or figures. Always explain in detail what the tables or figures show and what the reader expects to conclude from them. If you have a figure or table, the text should refer to it at least once.

- You can use suitable typesetting software that produces legible PDF output (LaTeX, Word, R Markdown, etc.). There is no strict page limit so you can use a readable font (e.g., 12 pt serif font), margins, and appropriately sized figures. Note that Jupyter Notebooks often lead to poorly formatted pdf. Out of curiosity, I took a random sample of 16 similar final reports that got total points from other courses I lectured. The task was identical to this one but without self-grading (which may add a page). The page counts of these final reports were 7, 7, 9, 10, 10, 10, 12, 12, 13, 13, 14, 14, 14, 14, 14, and 14. The reports had between 7 and 14 pages, the median being 12.5.

Even though you can modify your approach and adjust your algorithms for the final report, you are not required to (and probably should not) make significant changes. The idea is to polish the report and complete whatever steps you have planned.

The term project (final report and challenge submission) will be graded on an integer scale from 0 to 5 (1-5 = pass); see the grading criteria below.

The final report will be processed by the Ouriginal plagiarism detection system.

Grading of the term project

At the end of the course, you will be asked to give your project deliverables (final report, presentation, and challenge submission) an integer grade on a scale from 0 (fail) to 5 (excellent). You should attach the grading comments as the last section of your final report ("grading section"). The length of the grading section should be at most 1 page.

All group members will usually receive the same grade for this part of the course. (The group members may receive different grades if there are substantial problems with the contributions of some group members. Please get in touch with the course staff as soon as possible if there is any problem resolving them!) The course staff will consider this self-review when giving you the grade for the term project.

Grade for the deliverables

Please use the following grading guidelines to grade your group's deliverables (final report, presentation, and challenge submission) with a single integer grade from 0 to 5. **Please state the grade you gave yourself clearly at the beginning of the grading section!** Your deliverables may have shortcomings in one area, which better results in another area can compensate. You should try to balance weaknesses and strengths and produce one grade that faithfully describes your group's deliverables.

Notice about the challenge submission: the R2 score (and other performance measures) of the predictions on the test data is not a grading criterion by itself, even though a low R2 score may indicate that there is something else fishy in your approach which could affect grading.

In addition to the numeric grade, explain briefly (max. 1 page) the reasons for your grading using the grading criteria described below. The grading criteria are like the Data Science Master's Thesis assessment criteria. Please do not just repeat the grading criteria; tell how they apply and relate to your work.

Grade 5 (excellent): The treatment of the topics shows in-depth understanding, the relevant source material is used and cited, and the discussions show maturity. Appropriate machine learning and other methods have been chosen and applied correctly. The methods used have been analysed sufficiently. The reporting is to the point and exact. The conclusions drawn are in-depth and to the end. The discussion of findings shows an aptitude for independent, critical, and innovative research and thinking. The reports and presentations are polished and "camera-ready." The work has been creative and independent and progressed within the given schedule. The deliverables have been done by using the instructions provided.

Grade 3 (good): The treatment of the topic shows an understanding. The subject and literature are mainly analysed critically. The research material and methods (incl. machine learning methods) are suitable for the problem, and their use is well-argued. The findings have been reported in a primarily clear manner. The research questions are answered feasibly. The language is exact, and the terms used have been defined. The presentation is accurate, although the style may vary. The work has primarily proceeded to the planned timetable. The deliverables mostly follow the instructions given.

Grade 1 (passable): The topic and scope have not been motivated clearly, nor have the subject and goals fully understood. The work shows significant shortcomings in domain knowledge, and the cited sources are generally few or sub-quality. The reporting and analysis of the results have substantial weaknesses. Conclusions and discussions do not follow the scientific style. The deliverables are unpolished. The work has not progressed as planned. A substantial portion of the instructions given was not followed. However, the work still satisfies minimal requirements to be accepted.

Grade 0 (fail): The deliverables fail to satisfy the minimal requirements.

Grade for the group as a whole

Please also give your group a single integer grade from 1 to 5 and briefly (typically 1 paragraph of text) explain your grading. You can use the following rubric as a guideline, even though you do not need to grade each criterion separately. This grade does not directly affect the computation of your course grade. If you did the term project alone, you do not need to do this part.

Criteria	Grade: 5	Grade: 3	Grade: 1
Discussions about the content	The group has analytic and critical discussions. The discussion includes insights from the group members' own experiences. There is little irrelevant chatter.	The discussions are mainly about the topic of the project. There are examples from own experiences. Off-topic discussions are limited.	There are some discussions about the topic of the project. Some examples of own experiences are discussed, but they remain separate from the rest of the work. There are many off-topic discussions or discussions about topics of little relevance.
Setting the objectives and working towards the objectives	The group has a common goal which considers the individual objectives of the group members. The group works so that all the objectives are reached, and the objectives are – if necessary – adjusted during the progression of the work.	The group has a common objective that considers individual objectives to some extent. The group works towards objectives in an organised manner, even though all of the objectives may not be reached.	The group does not have a common objective. The group members work separately and do not share their responsibilities equally. A group has members who do not do their fair share of the work.
Participation, taking responsibility, interaction, atmosphere	Everyone participates actively in the discussions and group work. All group members take responsibility for the group work but also give room for the ideas of the others. Responsibilities are distributed fairly. The atmosphere of the group encourages to learn and do the work. Any conflict situations are resolved and learned from.	The group members participate in the meetings actively. The responsibilities and workload have been distributed fairly. The atmosphere is good, and an attempt is made to resolve conflicts.	The group has difficulties agreeing on meeting times, and all group members do not participate in the meetings. The distribution of responsibilities and work is uneven. Some group members do most of the work, while others do almost nothing. The atmosphere in the group does not encourage to learn in the group. The conflicts are not resolved.
Results and added benefits from the group work	The group work substantially contributes to the group members' learning outcomes.	The group advances the quality of the learning of its members to some degree.	The group brings no additional value to the learning of its members.