

# Report Group G1

- **Project title:** Thermophysical property: melting point (Kaggle). Prediction of melting points of organic molecules based on presence and number of functional groups
- **Team members:** Linus Krüer, Helena Windolf, Katharina Kirchmayr
- **Github repository:** [https://github.com/linuskruer/ids2025\\_group7\\_project\\_g1](https://github.com/linuskruer/ids2025_group7_project_g1)
- **Link to Kaggle competition:** <https://www.kaggle.com/competitions/melting-point/overview>

## 1.) Business understanding

### Identification of business goals

#### Background

The knowledge of the melting points of organic molecules is crucial in chemistry and chemical engineering. It enables compound identification and influences - among other things - process design, material selection and process safety. While the melting points of most common and widely used compounds are known, those of rare or newly synthesized ones remain obscure. The usual approach is experimental measurement of the melting points. Predictive models do exist, but are not able to reach the required accuracy and thus reliability yet, so experimental determination remains the gold standard so far. A subgroup of predictive systems for this task are based on so-called "group contribution methods", which estimate the properties of a chemical compound by subdividing its structure into smaller functional groups and summing the contributions of these groups.

#### Business goals

The overarching goal of this project is to develop a predictive system that estimates the melting points of organic molecules. Replacing laboratory measurements by mathematical predictions would accelerate processes and reduce costs. While the project is not conducted for a specific company or business, the development of such a system could benefit companies in the chemical and pharmaceutical sector as well as research institutes.

#### Business success criteria

The described overall business goal can be judged to be successful, if predictive models achieve an accuracy that is sufficient for most applications or at least for specific use cases. While this project alone will certainly not be able to reach this, it can make a small contribution to achieving this goal in the future.

## Assessment of situation

### Inventory of resources

- people: 3 students with knowledge in data science; 2 of them are students in the main study field Chemistry and Biochemistry
- hardware: personal laptops
- software: python3 and libraries; VS Code and Jupyter Notebook; Github
- data: training- and testing set from Kaggle competition

## **Requirements, assumptions, and constraints**

The project has to be completed by midday of 08.12.2025 due to the internal course deadline. While there is no concrete requirement regarding the performance of the resulting model, it will be tried to achieve one with a mean absolute error of less than 25%. Going forward in the project, we assume that the available data is correct and was acquired in a proper way. We further assume that the required software will stay available for the duration of the project and that our computers have the necessary computing power. The project is constrained by limited time availability. A minimum of 90 working hours will be spent on it, but possibly not considerably more than that. Furthermore, we are confined by our limited knowledge and experience in what kind of methods we will be able to apply, constraining the final result we will manage to achieve.

## **Risks and contingencies**

Factors that could potentially delay the completion of the project include Wifi outages, server problems and issues with the used software. We are prepared to tackle this by working from different devices or work spaces, keeping copies of all files locally on multiple devices and overall by planning completion of the project with enough time buffer.

## **Terminology**

- Tm: abbreviation for melting point in Kelvin (K)
- SMILES (Simplified Molecular Input Line Entry System): notation form of chemical structures using short strings
- functional group: group of atoms within a molecule that determines its chemical properties
- instance, element: rows of the data sets; corresponding to individual organic molecules
- features: counts of the functional groups
- label, target variable: value, that is tried to be predicted; in this case the melting point
- mean absolute error (MAE): average of the absolute differences between predicted and actual values
- K-Nearest Neighbours (KNN), Random Forest (RF), Support Vector Regression (SVR), Linear Regression, Neural Networks: machine learning algorithms
- hyperparameters: parameters that determine how a machine learning model learns
- randomized search, grid search: methods used to find the best hyperparameters for machine learning models by searching through combinations of hyperparameters

## **Costs and benefits**

There are neither costs nor financial benefits associated with the project. All used software is free of charge.

# **Definition of data-mining goals**

## **Data-mining goals**

Our first and foremost goal is to train a machine learning model for the prediction of melting points of organic molecules based on their composition of functional groups. Our objective is to achieve this by applying different methods on the available data and comparing their performance. We further want to explore the given data to find interesting structures or patterns and visualize these. All those results are aimed to be summarized and presented in the form of a poster. On a different note, another goal is to get hands-on experience in applying different machine learning algorithms and data science methods and deepening our understanding of those.

### **Data-mining success criteria**

The accuracy of our resulting model(s) will be assessed by their performance on a hold-out test set with the score of mean absolute error (MAE). Our personal goal is to reach a MAE of less than 25% and/or to end up in the top 100 of the underlying Kaggle competition. Due to the latter of the above stated goals, the project will also be judged successful even if this cannot be achieved, however.

## **2.) Data understanding**

### **Data gathering**

#### **Data requirements**

For our project, we require data that contains a sufficiently high number of different organic molecules, their melting points and their composition of functional groups. Ideally, the data should already be in a numerical format and be able to directly be loaded into python (or at least have the possibility to be converted into such a format). Not necessarily required for the machine learning itself, but for interpretation and exploratory data analysis, it would be beneficial to be able to map the instances to concrete organic molecules.

#### **Verification of data availability**

The availability of data containing the composition of functional groups and melting points of molecules could be verified. In the chosen data set, the identity of the individual molecules is depicted by so-called SMILES notation, while the different functional groups are simply denoted by a running number (Group 1, Group 2, Group 3, etc.). So far, no data could be found that maps those to specific functional groups (their chemical names, their SMILES notation or such). As the knowledge of the identity of the different groups is not necessary for training and predicting, this would rather only be a "nice to have" and not pivotal, however. If desired for interpretation purposes, the identity of individual groups that influence the melting point most could also be manually determined by checking the structures of several molecules containing that respective group.

#### **Data selection and collection**

All the data used to address the question comes from the Kaggle competition "Thermophysical Property: Melting Point". It is composed of 2 csv files (train set, test set), as well as an example submission file (also in csv format). As the data was presumably already selected from primary sources to serve the purposes of the competition, there are no dispensable columns or rows, so it will be used in its entirety. The possibility to access the data was already tested and verified: It could be loaded into python via the pandas library. The data had the expected size and format.

## Description of data

As already stated, the used data for this project originates from the Kaggle competition "Thermophysical property: melting point" and is composed of 2 files (train- and test set) in csv-format. The training set contains 80% of instances, which is a total of 2662 organic molecules. The test data contains the remaining 20%, which is equal to 666 instances. The columns of the data sets are id number, the identity of the molecule in SMILES notation and 424 different functional groups. The latter are (as outlined above) only denoted by numbers, not revealing what kind of groups they correspond to in chemical terms. The values of those columns are the amounts with which the individual groups occur in each molecule. In the train data set, an additional column is present that contains the melting points. The data is generally very sparse: most of the molecules contain only very few different groups. Also, some of the groups occur in a lot of the molecules, while the majority of groups are contained in almost none of the molecules. 87 columns of the train set even contain only zeros, meaning that the corresponding group is contained in none of the molecules. The data is sufficient for performing the planned methods and analysis tasks.

## Data exploration

As expected, the values of id numbers and SMILES names are unique. With regards to the melting points, the following could be observed: The minimum melting point lies at 53.54 K, the maximum one at 897.15 K. Their distribution is bimodal, with peaks at around 200 and 300 K. The mean lies at 278.26, with a standard deviation of 85.12. The majority of molecules in the data set have a melting point between 150 and 400 K. All distributions of the counts of the different functional groups are extremely right-skewed. The maximum by far lies always at 0 counts. 87 columns of the train set and 179 columns of the test set even contain only zeros, meaning that the corresponding group is contained in none of the molecules. 259 columns of the train set contain only zeros or ones (respective group is either not contained or only once contained in any molecule). There are no missing values in either set.

## Verification of data quality

No obvious quality issues could be found with regards to the data: It is complete, contains no missing values and has suitable format. It was checked that there are no molecules, that have no functional groups assigned (0 in all columns). The correctness of the data/whether it generally makes sense was verified in case of a few random samples. Without knowing which functional chemical groups correspond to which columns, this is not trivial, however. As the data stems from a reliable source, it is expected to be correct, however.

## 3.) Project plan

### 1.) Creation and set-up of repository on Github

- creation of repository
- connection of repository to local folders
- learning how to use Github (2 team members have not used it before)
- time per team member: 3 h

## **2.) Machine learning, part 1: model training**

- training different models on the training data
- different algorithms/methods (that were covered in the course) used: K- Nearest Neighbours (KNN) algorithm, Random Forest (RF), Support Vector Regression (SVR), Linear Regression, Neural Networks
- potentially further methods (that were not covered in the course) explored
- will be implemented in python3 (using, among others, the skikit-learn library) in the code editors VS Code or Jupyter Notebook
- time per team member: around 3 h
- individual methods divided between the team members:  
Katharina: Random Forest & SVR  
Helena: KNN & Neural Networks  
Linus: Linear Regression & potentially further methods

## **3.) Machine learning, part 2: hyperparameter tuning**

- attempting optimization of the hyperparameters for each method
- methods used to achieve this will include randomized search and grid search (both in combination with cross validation), Optuna and possibly further optimization methods
- submissions to Kaggle in-between to check scores on the test set
- used programs like in task 2
- time per team member: around 12 h
- division of methods between team members same as for task 2

## **4.) Visualization of data**

- exploration of different ways to visualize the data and to find interesting structures and patterns; only most interesting/meaningful results will be included in the end
- methods planned to be tried out: plotting, computation of correlations between individual functional groups and the melting point, clustering
- implemented in python3 using the editors VS Code or Jupyter Notebook; used libraries: Matplotlib, skikit-learn,...
- time per team member: around 4 h

## **5.) Creation of poster**

- selection and optimization of figures; writing of accompanying texts; layout & formatting
- will be done using LaTeX in Overleaf
- time per team member: around 10 h