Haute école d'ingénierie et d'architecture Fribourg
Hochschule für Technik und Architektur Freiburg

PS5 - Prédiction des propriétés chimiques par machine learning
Cahier des charges

Informatique et Système de Communication (ISC), 2022-2023

Etudiant	Simon Barras
Enseignants responsables	Prof. Beat Wolf - HEIA-FR Assis. Jonathan Donzallaz - HEIA-FR
Mendant	Inst. ChemTech Prof. Roger Marti - HEIA-FR Prof. Florence Yerly - HEIA-FR

v1.0
Fribourg, HEIA-FR, 14 octobre 2022

Historique des versions

Version	Changements	Date
0.1	Document créer sur Microsoft Word avec les chapitres contexte et objectifs	05.09.2022
0.2	Migration du document sur LaTeX et rajout de la première version du chapitre activités	9.10.2022
0.3	Finalisation du premier rendu du rapport	10.10.2022
1	Version finale du cahier des charges. Les problèmes de pages blanches et de glossaire ont été réglés. De plus, le planning a été mis à jour	14.10.2022

Table des matières

1 Contexte

Les liquides ioniques sont des éléments qui sont beaucoup utilisés en chimie. Ils permettent notamment de stocker de l'énergie mais disposent aussi de nombreuses autres applications.

Les liaisons ioniques s'obtiennent par l'attraction de deux ions de charge opposée. Les ions de charges positives et négatives sont respectivement appelés : cation et anion. Généralement, les cations sont métalliques tandis que les anions ne le sont pas. Un ion est composé d'un ou plusieurs atomes chargés électriquement.

Les possibilités de liaisons sont probablement infinies et leurs propriétés diffèrent pour chaque composition. Dans notre cas, la particularité qui nous intéresse est le point d'ébullition. Si on voulait obtenir cette information par expérience, il faudrait compter en jour le temps et les ressources consommées seraient importantes. L'objectif de ce travail de semestre est d'estimer la température d'ébullition d'un liquide ionique afin de concentrer le temps et les ressources de l'institut ChemTech de la HEIA-FR sur des liaisons validées par l'outil au préalable.

Ce travail sera utilisé comme outil de base pour un projet visant à stocker ou récupérer de l'énergie dans les phases de changements d'états d'un liquide ionique. Le but est d'y insérer en entrée le Simplified Molecular Input Line Entry Specification (SMILES) de l'anion et du cation et on obtiendra la température estimée.

L'écriture SMILES est utilisée en chimie pour décrire des molécules. Cette écriture permet de reconstruire le modèle 3D ou 2D comme ci-dessous :

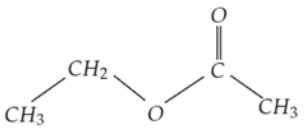

Figure 1.1 – Example de SMILES

Il existe une librairie python qui permet de récupérer les informations issues d'un SMILES.

Ce projet a déjà été amorcé par Mme. Yerly et un étudiant faisant un master de mathématiques à la HEIA-FR. Ils ont réussi à reproduire les résultats obtenus dans une étude à l'aide de la base de données fournie par cette dernière et vérifier par la base de données de l'école. Ce modèle était basé sur l'algorithme SVM (Support Vector Machine).

Ces algorithmes ont besoin de données d'entrées et de cibles. Son but sera de faire le lien entre les différents paramètres insérés et la cible la plus probable.

Dans un second temps, on pourrait aussi imaginer améliorer l'outil en rajoutant une surcouche qui déterminera le meilleur liquide ionique pour un objectif donné ou la possibilité de prédire d'autres propriétés.

2 | Objectifs

Les objectifs listent les différents délivrables du projet et les principaux se retrouveront retrouves dans la liste des milestones.

2.1 Principaux

— Créer une base de données regroupant les sources de données déjà utilisées et normalisées pour l'apprentissage

— Reproduire les résultats du projet de M. Yerly en utilisant les mêmes technologies

— Créer un nouveau modèle en utilisant les données extraites des SMILES

2.2 Secondaires

— Définir l'algorithme de machine Learning le plus efficace pour cette tâche

— Améliorer la base de données

— Améliorer l'expérience utilisateur

3 | Activités

Les activités sont toutes les tâches à effectuer pour réaliser le projet. Ces tâches viennent directement des objectifs et elles sont organisées en 3 niveaux afin de faciliter l'intégration à Gitlab.

3.1 Cahier des charges

Le cahier des charges doit être rédigé par l'étudiant et validé par tous les participants du projet. Il sera aussi défendu devant la classe afin de présenter le projet à tout le monde.

Chapitre 1 : Contexte

Must be done :

— Ecrire le chapitre contexte [2]

Chapitre 2 : Objectifs

Must be done :

— Ecrire le chapitre objectifs [1]

Chapitre 3 : Activités

Must be done :

— Ecrire le chapitre activités [2]

Chapitre 4 : Planning

Must be done :

— Ecrire le chapitre planning [1]

— Créer le diagramme de Gantt [3]

Défense du cachier des charges

Must be done :

— Utiliser et adapter un template latex pour le cahier des charges [2]

— Publier le cahier des charges sur le gitlab [2]

— Créer une présentation [3]

3.2 Reproduction du modèle SVM

Le modèle SVM est un modèle de machine learning qui a été utilisé dans l'article. Le reproduire permettra de poser les bases pour la suite et sera utilisé comme référence pour les prochains modèles.

Prendre en main le modèle SVM

Must be done :

— Prendre les données en main [1]

— Expérimenter SVM [1]

Créer un modèle avec les SMILES

Must be done :

— Utiliser l'algorithme SVM [5]

— Comparer les résultats avec le projet de Mme. Yerly [3]

— Documenter la démarche de reproduction [2]

3.3 Base de données normalisées

Les données sont les fondations des modèles de machine learning, c'est pourquoi cet objectif sera le premier à être réalisé. Il est important de bien structurer les données afin de pouvoir les utiliser facilement pour les futurs développements.

Analyser les données

Must be done :

— Identifier les données de l'article et de l'école [1]

— Repérer les données divergentes à l'aide du modèle [2]

— Documenter les sources de données [1]

Can be done :

— Identifier d'autres sources de données [1]

— Tester les autres sources de données [3]

Normaliser les données

Must be done :

— Définir les règles de normalisation [1]

— Définir la méthode de publication des données [1]

— Publier les données normalisées [2]

— Documenter l'infrasctucture et la normalisation [1]

3.4 Modèle avec les smiles

L'écriture SMILES encapsule les informations d'une liaison chimique. Afin de créer des modèles plus performants, on a besoin de ces informations que nous pouvons extraire. Plus tard, on pourra utiliser un deuxième algorithme ou améliorer l'expérience utilisateur en CLI ou avec une interface graphique.

Créer un modèle de machine learning

Must be done :

— Extraire les données des SMILES [3]

— Utiliser l'algorithme SVM avec les SMILES [8]

— Comparer les résultats avec la version précédente [2]

— Doucmenter l'extraction des données et le nouveau modèle [1]

Tester d'autres algorithmes

Can be done :

— Essayer le deep learning [3]

— Essayer le clustering [3]

— ...

— Documenter les résultats [1]

Améliorer l'UX

Can be done :

— Créer une interface graphique [3]

— Créer une API [3]

— Améliorer le CLI [3]

— Créer une documentation pour les outils effectués [1]

3.5 Défense du projet

La défense est la présentation du projet qui intervient une semaine après la remise du rapport.

Présentation du projet

Must be done :

— Créer une présentation [4]

4 Planning

Pour faire le planning, j'ai décidé de l'intégrer le plus possible à Gitlab. En effet, j'ai choisi d'utiliser les épics, les issues et les milestones disponibles directement sur Gitlab. Les avantages, c'est que le planing est très bien intégré et qu'on peut récupérer l'historique précis des actions effectuées pour compléter la tâche. En revanche, le diagramme de Gantt n'est pas pratique à utiliser et le point de vue est orienté sur les tâches et non sur le timing.

Pour organiser le travail, j'ai choisi d'utiliser deux niveaux d'épics. Le premier représente les objectifs principaux, ceux-ci sont complétés par des sous épics qui sont les tâches à effectuer. Ce deuxième niveau est complété par des issues qui sont les étapes à cocher pour compléter une tâche. Sur ces issues, on indiquera le poid afin de pouvoir quantifier l'effort nécessaire à la réalisation et on pourra lier des merge requests pour garder une trace des modifications. Les milestones sont là pour rappeler les échances de différents rendus.

le planning est directement accessible dans le menu "Roadmap" du groupe Gitlab (lien).

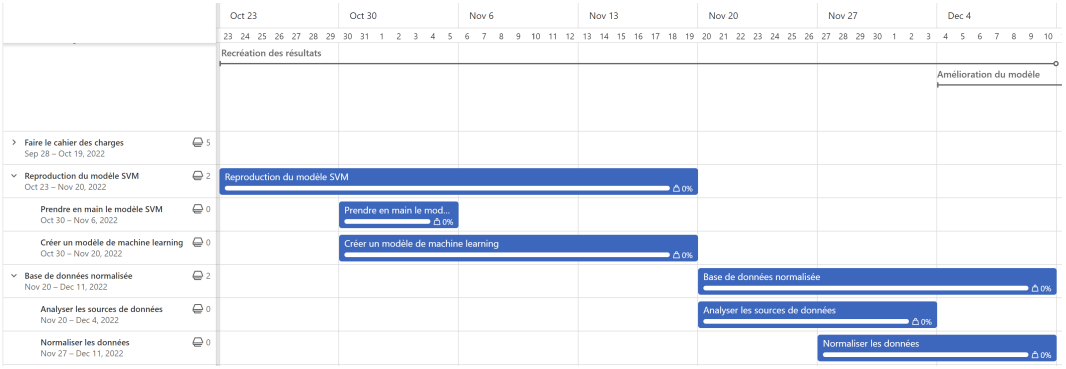

Figure 4.1 – Planning de la recréation des résultats

Ce premier planning représente la milestone qui correspond à la recréation des résultats. Il est composé des 2 épics principaux qui sont la base de données et le développement du premier modèle d'intelligence artificielle.

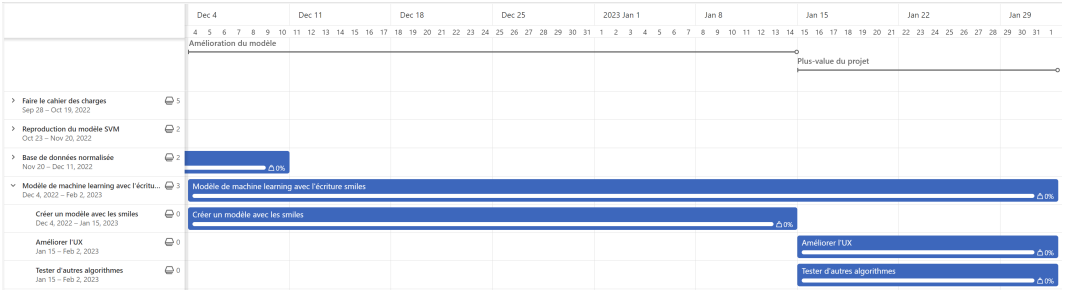

Figure 4.2 – Planning de l'amélioration du modèle

Cette deuxième partie de planing met en avant l'amélioration du modèle de machine learning à l'aide de l'écriture SMILES. Elle est coupée par une milestone qui correspoond à la décision qui sera prise pour la finalisation du projet. En cas de résultats satisfaisants, nous nous concentrons sur les améliorations de l'expérience utilisateur avec une API, une interface graphique ou autre solution. Dans le cas contraire, nous pourrons essayer d'autres algorithmes afin d'améiorer les résultats de la prédiction.

Table des figures

Acronymes

API Application Programming Interface. 7, 8

SMILES Simplified Molecular Input Line Entry Specification. 1, 2, 4, 5, 7, 8