Tím 20 - FOTOTOX

Retrospektíva šprintu č. 2

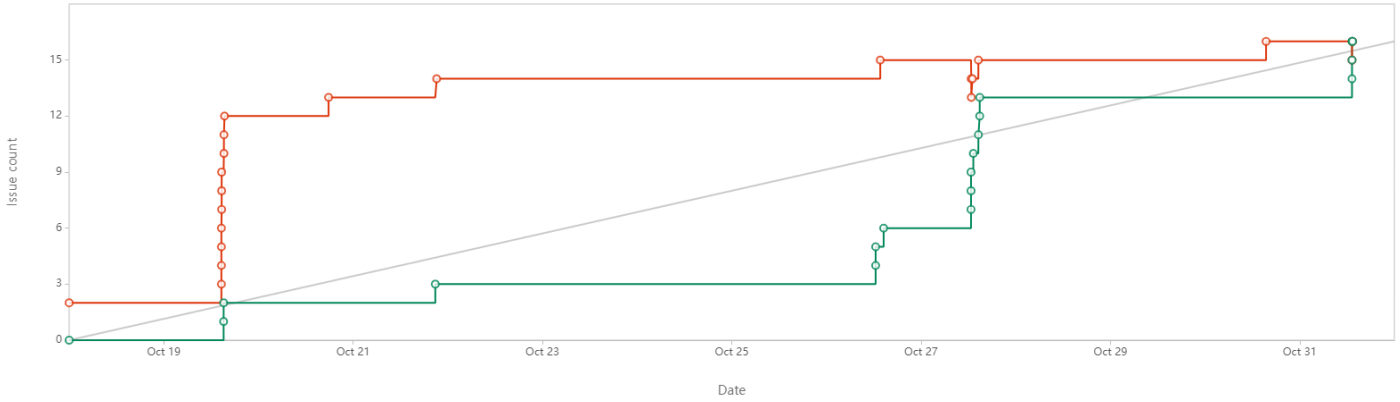
Začiatok šprintu: 18.10.2021
Koniec šprintu: 31.10.2021

Sumarizácia šprintu

Cieľom šprintu bola implementácia rozvinutých modelov pre predikciu a vyriešenie problému získavania nových dát. Podarilo sa nám implementovať nové systémy, ktoré nám do budúcnosti ušetria viac času. Pokročili sme vo všetkých smeroch, či už v rámci riadenia projektu, ale aj samotného vývoja - data engineering a modelling. Začali sme tiež plánovať štruktúru finálnej webovej aplikácie.

Zároveň sme narazili na niekoľko problémov, primárne stále zápasíme s dátovou doménou, či už s množstvom dát, prístupu k nim, alebo vhodnosťou rôznych deskriptorov.

Burnup graf



Export úloh

Issue Type	Issue key	Issue id	Summary	Assignee	Reporter	Status	Resolution	Created	Resolved
Task	PP-64	10063	Jira methodology design	Matej Halinkovič	Matej Halinkovič	Done	Done	31/10/2021 12:16	31/10/2021 12:16
Task	PP-56	10055	Cleaner clustering	Kateřina Mušková	Matej Halinkovič	Done	Done	27/10/2021 14:31	31/10/2021 12:16
Task	PP-51	10050	Populate CSV with Chem. properties (pipeline step)	Tibor Sloboda	Tibor Sloboda	Done	Done	21/10/2021 21:11	26/10/2021 12:29
Task	PP-50	10049	Data pipeline architecture	Matej Halinkovič	Matej Halinkovič	Done	Done	20/10/2021 17:46	27/10/2021 14:52
Task	PP-47	10046	CSV Merger	Tibor Sloboda	Matej Halinkovič	Done	Done	19/10/2021 15:14	26/10/2021 12:29
Task	PP-46	10045	Identify cyclic and acyclic compounds	Tibor Sloboda	Matej Halinkovič	Done	Done	19/10/2021 15:14	21/10/2021 20:50
Task	PP-40	10039	Ensemble learning exploration	Matej Halinkovič	Matej Halinkovič	Done	Done	13/10/2021 22:35	27/10/2021 14:32
Task	PP-39	10038	Deeper EDA - properly identify outliers, distributions, etc.	Kateřina Mušková	Matej Halinkovič	Done	Done	13/10/2021 22:34	27/10/2021 14:52
Task	PP-38	10037	Explore different architectures: Logit, SVM, ...	Matej Halinkovič	Matej Halinkovič	Done	Done	13/10/2021 22:33	27/10/2021 12:40
Task	PP-37	10036	Iteratively improve website	Patrik	Tibor Sloboda	Done	Done	13/10/2021 18:37	26/10/2021 14:30
Task	PP-32	10031	Re-train RandomForest with new data	Matej Halinkovič	Tibor Sloboda	Done	Done	13/10/2021 15:39	27/10/2021 12:40

Task	PP-29	10028	Meeting report 4	Jakub Maruniak	Tibor Sloboda	Done	Done	13/10/2021 15:37	27/10/2021 13:15
Task	PP-27	10026	Meeting report 3	Matej Halinkovič	Tibor Sloboda	Done	Done	13/10/2021 15:35	27/10/2021 12:40
Task	PP-21	10020	Obtain additional data	Jakub Maruniak	Matej Halinkovič	Done	Done	13/10/2021 15:23	19/10/2021 15:10
Task	PP-20	10019	TP Cup application	Jakub Knánik	Matej Halinkovič	Done	Done	10/10/2021 10:10	31/10/2021 12:15
Task	PP-9	10008	Meeting report 2	Jakub Maruniak	Tibor Sloboda	Done	Done	06/10/2021 14:06	19/10/2021 15:10

Čo sme spravili dobre?

- Implementovali sme nové postupy pre prácu s Jira, ktoré nám umožňujú lepšie spolupracovať na úlohách a udržiavať tak lepší prehľad o tom čo robia jednotliví členovia tímu
- Navrhli sme pipeline pre náš proces spracovania dát, ktorá nám umožňuje automatizovať väčšinu práce pre získanie nových informácií a teda nám v budúcnosti ušetrí veľa času
- Podarilo sa nám nájsť spôsob ako vypočítať nové relevantné deskriptory, ktoré neboli dostupné v databázach, s ktorými pracujeme
- Prihlásili sme sa do TP Cupu.

Čo nám robilo problémy?

- Stále sa nám nepodarilo vyriešiť nedostatok dát, najmä pre nefototoxické látky
    - vyplýva z toho problematické vyhodnotenie úspešnosti modelu. Náš prístup je úspešný na našom datasete, ale nevieme spoľahlivo predpovedať jeho úspešnosť v reálnom svete
- Zistili sme, že žiadny z členov tímu nemá zatiaľ skúsnosti s deploymentom modelov pre web aplikácie.
- Crawl dát z verejných stránok spôsobil, že sme dostali ban na zopár IP adries pre isté webstránky

Návrhy na zlepšenie

- Zistiť či sa dajú vypočítať všetky potrebné deskriptory zo SMILES kódu, ak áno, tak implementovať systém, ktorý to dokáže zvládať v reálnom čase
    - znížime tak závislosti na rôznych verejných databázach, ktoré by bolo problematické používať hlavne vo finálnom produkte
- Lepšie porozumieť chemickému pozadiu a princípom, ktoré spôsobujú fototoxicitu, aby sme mohli vytvoriť robustnejší a spoľahlivejší model
- Vyriešiť problém s nedostatkom dát
- Doučiť sa prístupy v web deploymentu modelov a začať pracovať na aplikačnej stránke nášho projektu