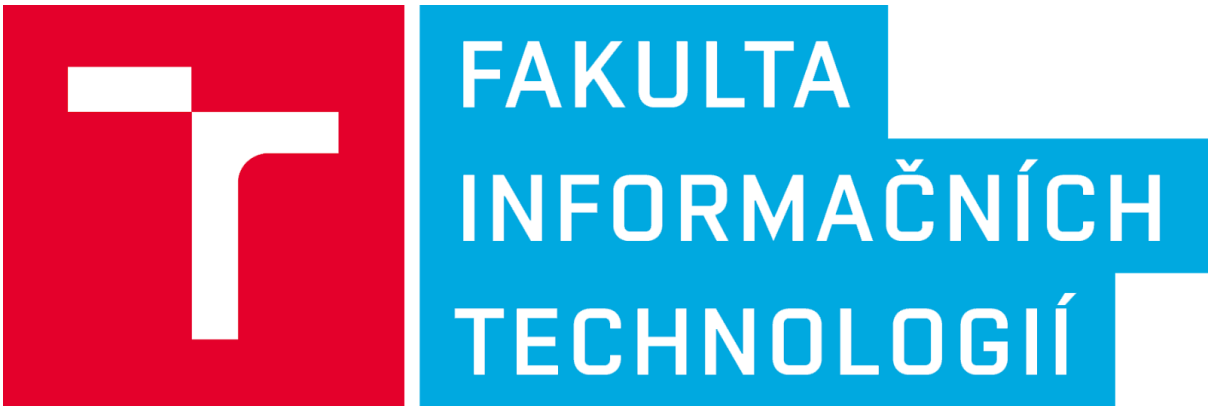


ZZN – Řešení

3. 1. 2018

A. Kvapilová (xkvapi12), T. Coufal (xcoufa09)

Dolovací úloha

Míra znečištění v závislosti na teplotě a vlhkosti

Úloha předpokládá závislost míry znečištění na okolních podmínkách a zkoumá trendy, které změny počasí provází. Na základě vysledovaných závislostí lze určit míru rizik spojených především s denní dobou, teplotou a vlhkostí. Míra znečištění je určena na základě evropských norem pro kvalitu ovzduší.

Rozbor úlohy

Určení míry znečištění je definováno pomocí tzv. AQI indexu. Ten je v různých zemích počítán jinak. V Evropské unii je tento index zjišťován z šesti různých faktorů[1]:

- 8h průměr koncentrace CO
- 8h průměr koncentrace O_3
- aktuální hodinová koncentrace NO_2
- aktuální denní koncentrace $PT_{2,5}$
- aktuální denní koncentrace PT_{10}
- aktuální denní koncentrace SO_2

V daném datasetu byly ovšem k dispozici pouze první tři zmiňované. Naopak jsou k dispozici hodnoty [2] koncentrací těžkých kovů, benzenu a NMHC a dalších, pro které ovšem existují pouze roční bezpečné limity – neexistuje specifikace pro určení kvality ovzduší na základě jejich momentálních hodnot. Z toho důvodu nebyly při výpočtu uvažovány.
Takto získaný AQI index definuje kvalitu ovzduší a míru znečištění jako kategorie: **Good**, **Moderate**, **Unhealthy**, **Very Unhealthy**, **Hazardous**, **Very Hazardous**.[3]

V řešené úloze jsme tuto kategorii znečištění predikovali pouze na základě teploty, absolutní vlhkosti, relativní vlhkosti, data a času, tedy pro takový systém, který by nemusel nutně měřit koncentrace látek v ovzduší.

Schéma řešení

1. Fltrace a předzpracování dat odstraněním irelevantních atributů a chybějících hodnot
2. Výpočet 8h půměrů pro CO a O_3 (*Moving Average* z rozšíření *Value Series*)
3. Klasifikace do AQI kategorií podle evropských norem (*Generate Attributes*, *Map*, *Set Role*)
4. Výpočet dílčích klasifikátorů, pro lepší predikci (*Generate Attributes*)
5. Klasifikace znečištění pomocí Naive Bayes (*Validation*, *Select Attributes*, *Naive Bayes*, *Apply Model*, *Performance*)

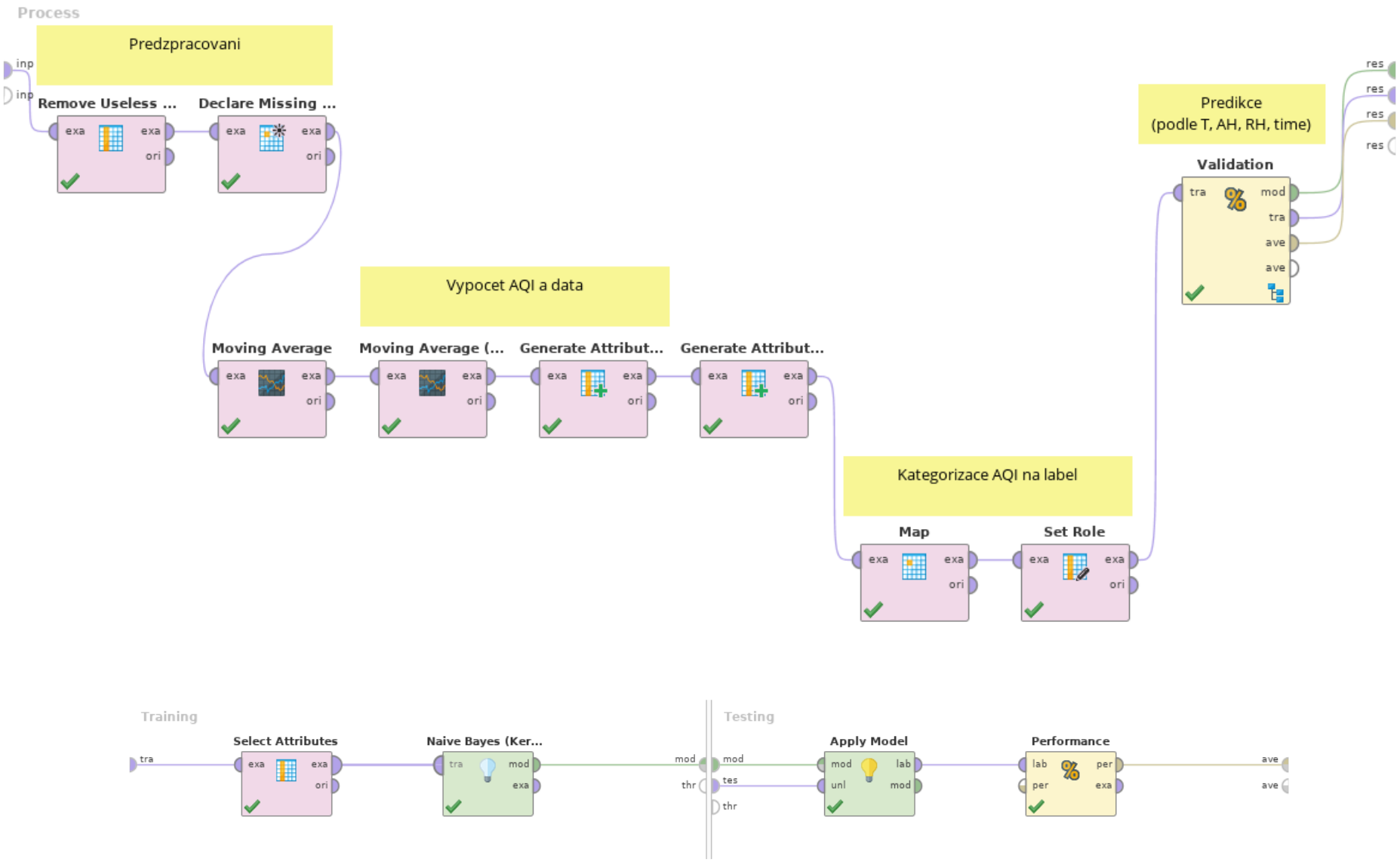
Pro lepší výsledky predikce jsme vedle zřejmých charakteristik jako teplota a vlhkost přidali klasifikátory pro:

- Den v týdnu
- Denní doba (noc, siesta, práce)
- Roční období
- Pracovní dny (resp. dny s charakteristicky nižší produkcí škodlivin: Sobota, Neděle a Pondělí)
- Hodina měření

Pro naučení prediktivního modelu jsme použili Naive Bayes klasifikaci, která je vhodná pro menší soubory dat a navíc dobře klasifikuje numerické hodnoty.

---

[1] https://www.airqualitynow.eu/about_indices_definition.php
[2] https://archive.ics.uci.edu/ml/datasets/Air+quality
[3] http://www.haze.gov.sg/docs/default-source/computation-of-the-pollutant-standards-index-(psi).pdf


Process
Predzpracovani
inp
inp
Remove Useless ...
Declare Missing ...
exa
ori
Vypocet AQI a data
Moving Average
Moving Average (...
Generate Attribut...
Generate Attribut...
Kategorizace AQI na label
Map
Set Role
Predikce
(podle T, AH, RH, time)
Validation
tra
mod
ave
res
Training
Testing
Select Attributes
Naive Bayes (Ker...
Apply Model
Performance
lab
per
unl
tes
thr
ave

Závěr

Výsledný model predikuje míru znečištění s 60% úspěšností. To je způsobeno především nízkou kvalitou vstupních dat:

- chybějící faktory znečištění
- vychýlením dat směrem k vysokému znečištění
- mnoho chybějících hodnot
- malý soubor dat (1 rok pozorování)
- Nedostatek meteorologických údajů (např. tlak)

I přesto byla klasifikace relativně úspěšná a pokud nebyl testovací vzorek zařazen do správné kategorie, byl modelem přiřazen do vedlejší (tzn. Pokud měl být vzorek Hazardous byl chybně zařazen jako Very Hazardous), nikdy však nedošlo k jeho zařazení do opačné části spektra znečištění.

Predikovaná klasifikace	**Správná klasifikace**			**Class precision**
	Very Unhealthy	**Hazardous**	**Very Hazardous**	
Very Unhealthy	600	173	253	58.48%
Hazardous	124	92	67	32.51%
Very Hazardous	278	184	892	65.88%
Class recall	59.88%	20.49%	73.60%	