
# VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

## FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

## ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING


# KONCEPT DRIFT V METABOLOMICKÉ ANALÝZE

CONCEPT DRIFT IN METABOLOMIC ANALYSIS

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS


AUTOR PRÁCE

AUTHOR

Aleš Koštoval

VEDOUCÍ PRÁCE

SUPERVISOR

Ing. et Ing. Jana Schwarzerová, MSc



BRNO 2022

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

# Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

***Student:*** Aleš Koštoval ***ID:*** 221515

***Ročník:*** 3 ***Akademický rok:*** 2021/22

**NÁZEV TÉMATU:**

## Koncept drift v metabolomické analýze

**POKYNY PRO VYPRACOVÁNÍ:**

1) Seznamte se s problematikou konceptu drift a jeho vlivu ve strojovém učení. 2) Prostudujte techniky používané pro detekci a korekci konceptu drift využívané v běžné praxi. 3) Otestujte minimálně dvě metody pro detekci a korekci drift konceptu na vhodném datasetu a srovnejte jejich výsledky. 4) Vyberte minimálně jednu otestovanou metodu a vhodně jí implementujte na metabolomický dataset. 5) Detekujte koncept drift v metabolomickém datasetu na různých predikčních modelech. 6) Proveďte diskusi k výsledkům.

**DOPORUČENÁ LITERATURA:**

[1] Tom Mitchell. Machine Learning. Maidenhead, U.K.: McGraw-Hill, 1997. isbn: 0-07-115467-1.

[2] Roberto Barros and Silas Santos. “An Overview and Comprehensive Com parison of Ensembles for Concept Drift”. In: Information Fusion 52 (Dec. 2019), pp. 213–244.

[3] Joao Gama et al. “Learning with Drift Detection”. In: Advances in Artificial Intelligence – SBIA 2004. Ed. by Ana L. C. Bazzan and Sofiane Labidi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 286–295. isbn: 978-3-540-28645-5

***Termín zadání:*** 7.2.2022 ***Termín odevzdání:*** 27.5.2022

***Vedoucí práce:*** Ing. et Ing. Jana Schwarzerová, MSc

**doc. Ing. Jana Kolářová, Ph.D.**
předseda rady studijního programu

**UPOZORNĚNÍ:**

Autor bakalářské práce nesmí při vytváření bakalářské práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

## ABSTRAKT

Tato bakalářská práce se zabývá problematikou strojového učení, konkrétně analýzou drift konceptu. Jedná se o nechtěný jev, který lze detekovat v predikčních modelech. Pomocí detekce s následnou korekcí drift konceptu se predikční modely stávají spolehlivějšími a jsou schopny adekvátně reagovat na vstupní data reprezentující dynamickou informaci. Za vhodného reprezentanta těchto dat lze považovat metabolomická data. Metabolomická data a jejich analýza může pomoc k včasné detekci nemocí jako je diabetes mellitus, či rakovina. V první části práce jsou popsány teoretické poznatky z oblasti analýzy drift konceptu a metabolomické analýzy. Druhá část pojednává o postupu modelování predikčních klasifikátorů a implementaci algoritmů pro detekci drift konceptu. Praktická část práce byla realizována v programovacím jazyce Python. Na závěr jsou v rámci druhé časti popsány dosažené výsledky a jejich diskuze.



## KLÍČOVÁ SLOVA

Koncept drift, Metabolomika, Predikční modelování, Strojové učení



## ABSTRACT

This bachelor thesis deals with machine learning, specifically the analysis of the concept drift. This is an unwanted phenomenon that can be detected in predictive models. Through detection followed by correction of the concept drift, predictive models become more reliable and can respond adequately to input data representing dynamic information. Metabolomic data can be considered a suitable representative of such data. Metabolomic data and their analysis can help to detect diseases such as diabetes mellitus or cancer early. In the first part of this bachelor thesis, the theoretical background of concept drift analysis and metabolomics analysis are described. The second part discusses the process of modeling predictive classifiers and implementing algorithms for concept drift detection. The practical part of the work was implemented in the Python programming language. Finally, the second part describes the results obtained and their discussion.



## KEYWORDS

Concept drift, Metabolomics, Predictive modeling, Machine learning



Vysázeno pomocí balíčku `thesis` verze 4.07; `http://latex.feec.vutbr.cz`

KOŠTOVAL, Aleš. *Koncept drift v metabolomické analýze*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2022, 55 s. Bakalářská práce. Vedoucí práce: Ing. et Ing. Jana Schwarzerová, MSc.

# Prohlášení autora o původnosti díla

| | |
|---|---|
| **Jméno a příjmení autora:** | Aleš Koštoval |
| **VUT ID autora:** | 221515 |
| **Typ práce:** | Bakalářská práce |
| **Akademický rok:** | 2021/22 |
| **Téma závěrečné práce:** | Koncept drift v metabolomické analýze |

Prohlašuji, že svou závěrečnou práci jsem vypracoval samostatně pod vedením vedoucí/ho závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ..... .....

podpis autora*

*Autor podepisuje pouze v tištěné verzi.

# PODĚKOVÁNÍ

Rád bych poděkoval vedoucí bakalářské práce paní Ing. et Ing. Janě Schwarzerové, MSc za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.

# Obsah

# Seznam obrázků

# Seznam tabulek

# Úvod

Bakalářská práce se zabývá problematikou drift konceptu v metabolomické analýze. Analýza drift konceptu spadá do vědního oboru strojového učení a pro popis dat využívá statistické vztahy. Pomocí těchto vztahů lze detekovat nežádoucí změny v distribuci dat, tzv. koncept drift. Tento nechtěný jev lze pozorovat na predikčních modelech, když na vstup modelu dorazí dynamicky se měnící data. Poté je rozhodování modelu tímto faktorem značně ovlivněno a tím pádem vznikají chybné predikce. Díky detekci s následnou korekcí změn v konceptu je možno zajistit dlouhodobou přesnost, spolehlivost a stabilitu predikčních modelů.

Pro analýzu drift konceptu byly vytvořeny metabolomické modely predikující tři klasifikační problémy reprezentující fenotypovou informaci. Metabolity nejlépe popisují molekulární fenotyp, ale zároveň se jedná o biochemické substance charakteristické pro svou dynamickou proměnlivost v čase. Proto byl výše naznačený princip detekce drift konceptu aplikován na predikční metabolomické modely.

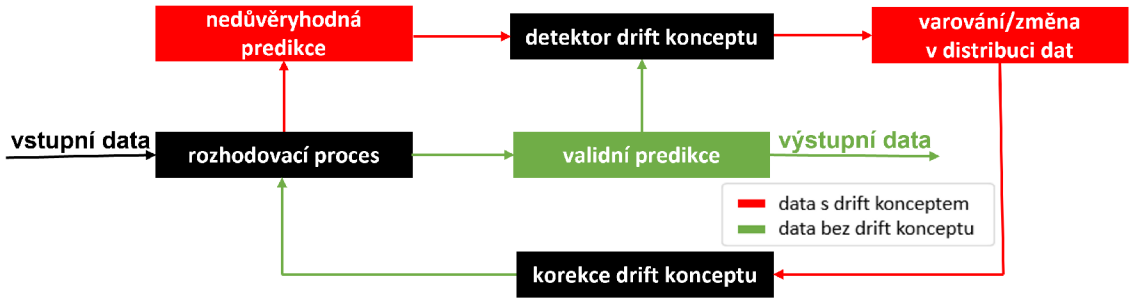
Práce je členěna do sedmi kapitol. První kapitola pojednává o drift konceptu a jeho zařazení v kontextu strojového učení. Nezbytná část teorie se zabývá novým vědním oborem studující metabolity, metabolomice. V další kapitole jsou zmíněny algoritmy použitých predikčních modelů. Metodologie detekce drift konceptu je popsána ve čtvrté kapitole. Následuje kapitola pojednávající o metodách korekce drift konceptu. Šestá kapitola popisuje použitá data. Praktická část je obsažena v sedmé kapitole, ve které jsou popsány použité predikční modely, na které byly aplikovány detektory drift konceptu. V návaznosti na použité metody byly popsány dosažené výsledky predikcí a detekce drift konceptu. Závěrem bakalářské práce jsou shrnuty její cíle a přínos.

Hlavní cíl bakalářské práce spočívá v implementaci algoritmů pro detekci drift konceptu na metabolomické predikční modely. Praktická část byla realizována v programovacím jazyce Python. Nejprve proběhlo učení predikčních modelů na trénovací množině dat, která představovala koncetrace metabolitů. Natrénované modely predikovaly tři klasifikační problémy představující fenotypovou informaci. Na výsledné predikce byly implementovány detektory drift konceptu pomocí knihovny Scikit-multiflow s cílem detekce změn v distribuci dat.

# 1 Drift koncept

Analýza drift konceptu náleží do vědního oboru strojového učení. Napomáhá zde odhalovat a korigovat nepřesnosti predikčních modelů. Tyto nepřesnosti jsou dány změnou v distribuci dat zapříčiněné dynamickou informací ve vstupních datech reprezentující reálný svět. Cílem detekce drift konceptu je navržení algoritmů, které budou v reálném čase tyto změny distribuce dat detekovat a následně revidovat, čímž se zajistí dlouhodobá přesnost, spolehlivost a stabilita predikčních modelů. [1]

Klasifikační predikční modely mohou reprezentovat rozhodovací proces, do kterého vstupují data s určitým pravděpodobnostním rozložením. Nicméně vstupní data se mohou kdykoliv nepředvídaným způsobem změnit. Ze statických dat přijdou na vstup data dynamická. Dynamická data značně snižují spolehlivost a přesnost rozhodovacího procesu, který není schopen na tuto změnu v distribuci dat zareagovat. Rozhodovací proces zareaguje, tzv. nedůvěryhodnou predikcí, čímž vznikají chybné predikce. Pro odhalení chybné predikce lze použít detektory drift konceptu, které nám odhalí případné změny v distribuci dat. Tyto chybné predikce jsou v druhém kroku vhodným korekčním algoritmem eliminovány a rozhodovacím procesem opět správně predikovány. Princip samotného drift konceptu je popsán na obrázku 1.1, kde vidíme výše popsaný princip analýzy drift konceptu. [2] [3] [4]




Obr. 1.1: Schéma principu drift konceptu

Další charakteristickou vlastností drift konceptu je frekvence změn v distribuci dat. Za optimální variantu považujeme, když se koncept mění plynule v čase. Někdy se však může jednat o náhlé změny, které se komplikovaněji detekují. Opakující se změny konceptu lze jednoduše detekovat, analyzovat a revidovat. [4]

Prvotní přístupy pro analýzu dynamicky se měnících vstupních dat byly založeny na následujících strategiích: aktivní učení v reálném čase, upřednostňování nedávných dat, nebo na dynamickém posuvném okně. Tyto přístupy kolabovaly v případě, kdy došlo k náhlé změně vstupních dat. Tím pádem klesala i přesnost samotného modelu, což bylo zapříčiněno nedostatkem metrik pro detekci náhlých změn v distribuci dat. [5]

Současná strategie pro analýzu drift konceptu ve strojovém učení vypadá následovně. Nejprve je potřeba vhodným způsobem drift koncept detekovat a zároveň monitorovat přesnost detekce algoritmu. Dalším krokem je použití statického modelu, který nám bude sloužit jako průběžný srovnávací ukazatel a bude signalizovat, zda se s časem model nestal méně přesným. Je zapotřebí, aby se tento statický model pravidelně přeučoval dle nových dat, což napomáhá tomu, aby byl model v souladu s dynamickými daty. Nová data mají přidělenou vyšší váhu na základě jejich aktuálnosti než data historická. Avšak se mohou vyskytovat i vstupní data podobná historickým, což je nutné zohlednit při implementaci. Proto je vhodné vytvořit nové modely, které nám odhalí náhlý nebo opakující se drift koncept, z důvodu nespolehlivosti statického modelu. Ten poté lépe detekuje plynulé změny v datech. Posledním krokem by byla samotná korekce drift konceptu, která signalizuje, zda byly tyto změny distribuce dat detekovány správně a koriguje detekované drift koncepty. [1]

## 1.1 Detekce drift konceptu

Detektory drift konceptu obecně analyzují výsledné predikce modelu. Většina algoritmů na detekci drift konceptu je založena na statistických parametrech a opírají se o matematické vzorce. Detektory aplikují rozhodovací pravidlo, na základě matematických vzorců, pomocí kterých detekují případné změny v distribuci dat. [1]

Detekční algoritmy signalizují, na modelem vytvořené predikci, drift koncept ve dvou prahových úrovní. Poklesne-li spolehlivost detekce pod první práh, dojde k upozornění, že by mohlo dojít ke změně v konceptu, tzv. varovná úroveň. Když dojde k překročení druhého prahu, je signalizována změna v distribuci dat, tj. úroveň změny. [6]

## 1.2 Korekce drift konceptu

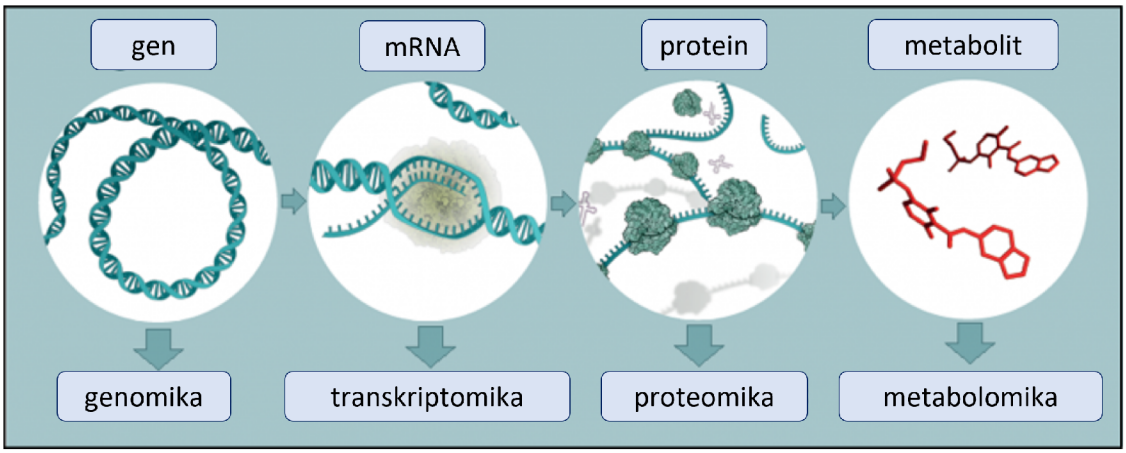
Algoritmy korekce slouží k úpravě detekovaných koncepčních driftů na základě ověření správnosti detekce, což eliminuje nesprávně označené změny distribuce dat. Většinou se jedná o algoritmy s detekcí a následnou korekcí drift konceptu.

Existují dva přístupy korekce. První z nich aktualizuje model ihned, když dojde k detekci drift konceptu a díky tomu je model schopen ihned reagovat na změnu v rozložení dat. Tyto algoritmy sledují distribuci dat v konkrétních úsecích datového toku. Druhý přístup spočívá v pravidelné aktualizaci modelu a to bez zřetele na to, zda došlo ke změně konceptu, či nikoliv. Jde například o metodu založenou na podpůrné vektorové regresi, o které blíže pojednává následující kapitola 5.2. [7]

# 2 Metabolomika

Metabolomika se zabývá detekcí, kvantifikací a analýzou metabolitů v daném organismu. Metabolity mají hlavní roli v rámci metabolických drah. Mohou být různé povahy. Jedná se o malé molekuly (do 1500 Da) proteinů, aminokyselin, organických kyselin, vitamínů, sacharidů atd. Metabolity hrají klíčovou roli pro pochopení fenotypů imunitních onemocnění. [8]

Metabolomika tak nedávno doplnila poslední chybějící článek do smyčky systémové biologie, která se skládá ze čtyř omických vědních oborů viz obrázek 2.1. První z nich je genomika zabývající se studiem genomu, tedy souborem veškeré genetické informace v daném organismu. Druhým je transkriptomika studující transkriptom, aneb soubor veškeré mRNA nacházející se v daném organismu. Proteomika se zabývá strukturou a funkcí proteinů. Metabolomika, studující metabolity, nejlépe popisuje molekulární fenotyp daného organismu. Jednotlivé zkoumané komponenty, uvedené zmíněnými vědními obory, jsou propojeny vzájemnými interakcemi.




Obr. 2.1: Zobrazení omických oborů. Převzato a upraveno z [9]

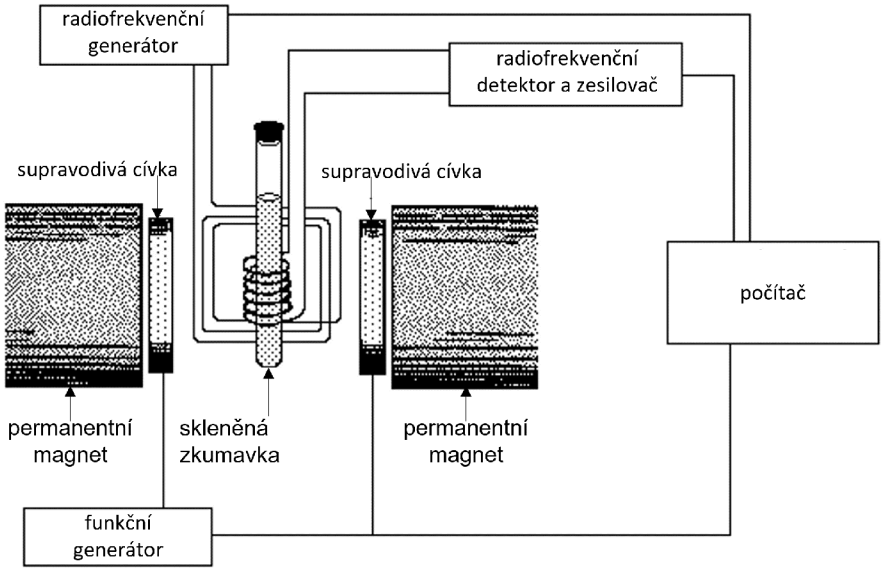
## 2.1 Analýza metabolomických dat

Jedním z prvních kroků metabolomické analýzy je meření dat. Data pro tuto práci byla získána pomocí spektroskopie nukleární megnetické rezonance (SNMR) a hmotnostní spektrometrie doby letu (TOFMS). Díky pokroku metabolických měřících technik lze prokázat řadu onemocnění, právě na základě nízkých molárních koncetrací některých metabolitů. Mezi nejznámější měřící techniky koncetrací metabolitů se řadí SNMR a TOFMS. SNMR využívá magnetické pole. Zatímco TOFMS je založena na dvou měřících principech - hmotnostní spektrometrie a doby letu. [10]

### Spektroskopie nukleární magnetické rezonance (SNMR)

SNMR je jednou z měřících technik metabolitů využívající vlastnosti molekul látek. Konkrétně na atomární úrovni jejich elektricky nabitých jader. SNMR je založena na principu působení vnějšího magnetického pole, kdy se jádra atomu molekuly látky nachází v různých spinových stavech mající určitý magnetický moment. Pomocí SNMR jsou detekovány pouze jádra s nenulovým spinem. Nelze tedy detekovat především izotopy, které mají sudý počet protonů i neutronů. [11]

Princip SNMR je znázorněn na obrázku 2.2. Na permanentní magnet je přivedeno konstantní magnetické pole o magnetické indukci 6-24 T generované supravodivými cívkami. Uvnitř permanentního manetu je umístěn biologický vzorek ve skleněné zkumavce. Zároveň jsou pomocí radiofrekvenčního generátoru na biologický vzorek aplikovány frekvence v rozsahu 60-100 MHz. Poté jsou radiofrekvenčním detektorem detekované frekvence emitované při změnách energie jader. Tyto emitované frekvence jsou zesíleny a detekovány počítačem. Výsledky jsou vizualizovány pomocí vhodného softwaru jako závislost chemického posunu na amplitudě píku. [11] [12]




Obr. 2.2: Schéma techniky NMR spektroskopie. Převzato a upraveno z [11]

### Hmotnostní spektrometrie doby letu (TOFMS)

TOFMS využívá dvou měřících principů, a to hmotnostní spektrometrie a doby letu. Obecně se hmotnostní spektrometrie používá ke stanovaní poměru hmotnosti iontů ku jejich náboji. Hmotnostní spektrometr se skládá z jednoho hmotnostního analyzátoru, nebo dvou hmotnostních analyzátorů. Toto spojení dvou analyzátorů

slouží pro zlepšení senzitivity a selektivity detekce a jedná se o hybridní hmotnostní spektrometrii. Princip detekce iontů je následující: přes první hmotnostní analyzátor jsou přenášeny primární ionty a po průchodu prvním analyzátorem se ionty vlivem ionizace rozpadnou. Ionizace může být způsobena světlem o konkrétní vlnové délce, např. laserem. Tyto sekundární fragmenty iontů se přenášejí přes druhý analyzátor a jsou detekovány iontovým detektorem. Hmotnostní spektrometry se liší typem použitého hmotnostního analyzátoru. [13]

Doba letu je jednou z variant, kterou lze použít jako hmotnostní analyzátor. Tento hmotnostní analyzátor dělí ionty na základě času potřebného na průnik letové trubice se známou délkou od iontového zdroje k detektoru, který se nachází na opačné straně trubice. Je zřejmé, že ionty s nižším poměrem hmotnosti ku náboji se budou pohybovat trubicí rychleji a budou detekovány jako první. Kinetická energie a hybnost následně ovlivňuje rychlost detekovaných iontů. Výsledkem bude grafická závislost poměru hmotnosti iontů na náboji a intenzitě detekovaných iontů. [13]

## 2.2 Predikční modely v metabolomice

Predikční modelování spočívá v učení modelů na rozsáhlých datech, které nelze pouhým lidským faktorem analyzovat. V principu jde o automatizaci rozhodovacích a prediktivních procesů vycházející ze statistické analýzy. Cílem těchto predikčních modelů je co nejdůvěryhodnější predikce nových vstupních dat. [14] [15]

Za vstupní data bývají pokládány příznaky daného objektu. Výstupní data jsou modelem přetransformovaná vstupní data do předpokládané podoby. Správně natrénovaný model by se měl učit z původních dat, na základě toho se adaptovat na změny a tím vytvářet predikci nových vstupních dat. Matematicky jde o aproximaci transformační funkce *f*, která je definována:

$$Y = f(X), \tag{2.1}$$

kde $X$ jsou vstupní data a $Y$ jsou data výstupní. Na základě vztahu 2.1 je zjišťována závislost mezi nimi, kterou vyjadřuje transformační funkce *f*.

Algoritmy strojového učení lze separovat do dvou kategorií dle způsobu učení algoritmu, a to na učení s učitelem a učení bez učitele. Učením rozumíme takovou schopnost algoritmu, pomocí níž je daný model schopen se přizpůsobit změnám vnějšího prostředí.

Algoritmy založené na učení s učitelem předpokládají, že ke vstupním datům jsou označeny cílová data, na kterých se může algoritmus natrénovat. Mezi algoritmy učení s učitelem se řadí metody klasifikace, kdy algoritmus na základě společných příznaků třídí objekty do již předem známých tříd. Další z algoritmů učení s učitelem je regrese, která zkoumá vztah mezi proměnnými (závislou a nezávislou). Lze tak

na základě známé nezávislé proměnné předpovědět, jak bude vypadat proměnná závislá. [14]

Algoritmy založené na učení bez učitele nepředpokládají znalost cílových dat. Nejčastěji používanou metodou je shlukování. Algoritmus si tedy musí sám najít podobné příznaky vstupních objektů a na základě toho klasifikovat objekty do shluků s podobnými příznaky. Většinou jde o rozsáhlejší a komplikovanější vstupní data než u předchozího přístupu učení s učitelem. [14]

# 3 Algoritmy predikčních modelů

V praktické části bakalářské práce byly použity algoritmy predikčních modelů založené na učení s učitelem. Ve všech případech se jednalo o klasifikační úlohy. Klasifikace proběhla pomocí algoritmů založených na logistické regresi, Naivním Bayesovském klasifikátoru, náhodném lesu a metodě gradient boosting.

## 3.1 Logistická regrese

Logistická regrese je metodou lineární. Jejím průběhem je sigmoidální funkce. Průběh sigmoidální funkce je možno popsat následně: v počáteční fázi exponenciálně roste a na konec se asymptoticky zpomaluje. Na základě počáteční podmínky, kdy sigmoidální funkce v 0 nabývá funkční hodnoty 0.5, se zvyšuje přesnost klasifikace. [16] [17]

Tato metoda bývá používána jako klasifikátor, který separuje objekty do předem známých tříd. V ojedinělých případech lze logistickou regresi použít k řešení regresivních úloh. Jedná se tedy o metodu používanou v přístupu učení s učitelem.

## 3.2 Naivní Bayesovský klasifikátor

Naivní Bayesovský klasifikátor s předpokladem normálního rozložení dat je založen na principu Bayesovy věty o podmíněné pravděpodobnosti. Za předpokladu nezávislosti normálního rozdělení atributů lze separovat jednotlivé příznaky vstupních dat. Dalším předpokladem pro použití této metody je normální rozdělení vstupních dat. [18]

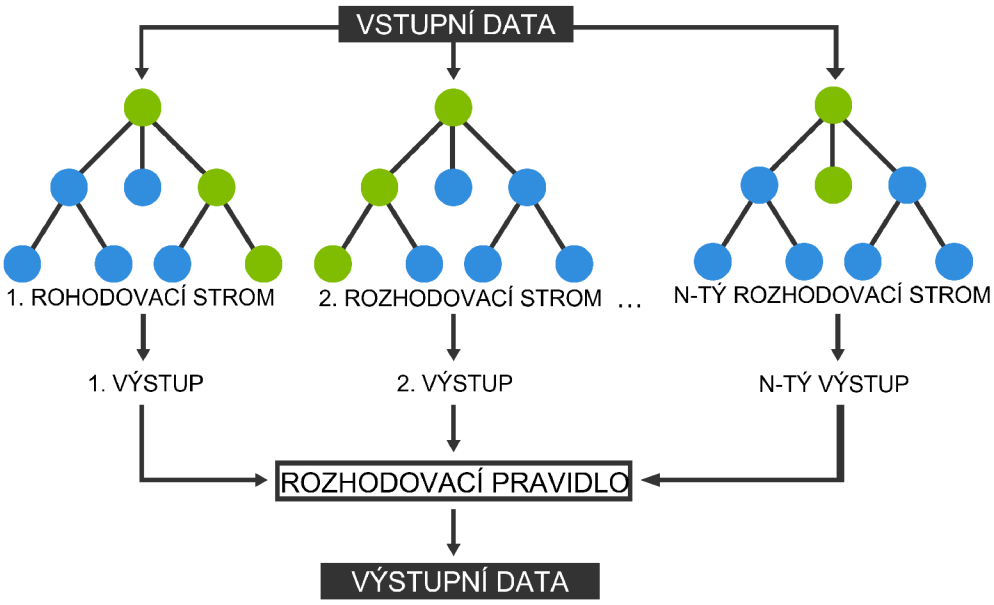
Vlivem paralelizace příznaků, klesá výpočetní náročnost, avšak i tento klasifikátor má určitá omezení. Tím je zmiňovaná podmíněná nezávislost příznaků, která může být i nevýhodou. Problém nastává tehdy, když data obsahují méně objektů než příznaků. Dále je vhodné, aby byl rozptyl v jednotlivých třídách stejný, čímž je splněna jedna z podmínek normálního rozdělení dat. [18]

## 3.3 Náhodný les

Náhodný les se používá jak pro klasifikační, tak i pro regresivní úlohy. Jde o relativně složitý algoritmus, který je založen na myšlence skládání rozhodovacích stromů a patří mezi souborové metody strojového učení.

Nejprve je nutno určit počáteční podmínky, a to z kolika stromů se les bude skládat a hloubku lesa. Princip algoritmu náhodných lesů lze popsat v následujících korcích viz obrázek 3.1 [19]:

1. Pomocí metody bootstrapping jsou náhodně vybrány některé příznaky ze vstupních dat reprezentující uzly rostoucího stromu. Tyto uzly se mohou definovaným způsobem spojovat s ostatními uzly.
2. Spojené uzly vytváří rozhodovací strom.
3. Dle definovaného počtu stromů jsou opakovány kroky 1 a 2.
4. Na základě rozhodovacího pravidla, např. průměru nebo modusu, je rozhodováno o zařazení objektu do náležité třídy. Přičemž nejvyšší váha rozhodování je udělena stromu, který měl nejvyšší průměrnou pravděpodobnost správného zařazení napříč všemi stromy.




Obr. 3.1: Schéma principu náhodných lesů. Zjednodušeně pro 2D data - zelený uzel reprezentuje jeden příznak a modrý uzel druhý příznak. Převzato a upraveno z [20]

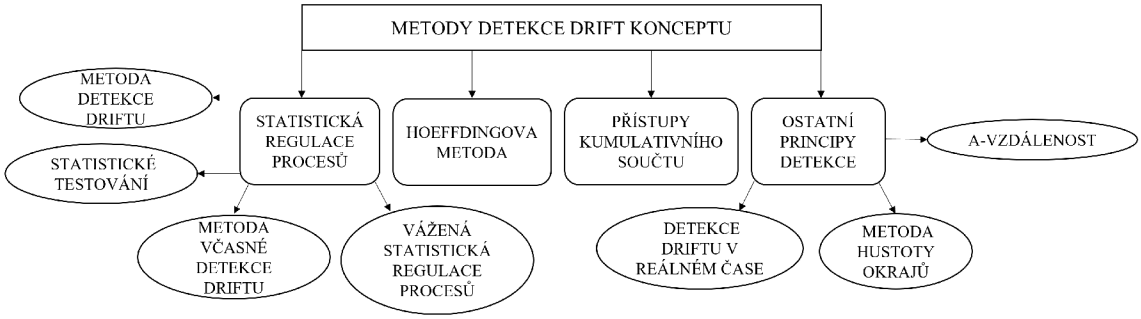
## 3.4 Gradient boosting

Metoda gradient boosting se řadí mezi souborové metody strojového učení a své uplatnění nachází v klasifikačních i v regresivních úlohách. Tato metoda je sofistikovanější verzí náhodných lesů využívající rozhodovacích stromů k zpřesňování predikce.

Rozhodování modelů založených na gradient boostingu vychází ze znalosti, že průměrováním predikcí více různých modelů dochází k značnějšímu zpřesnění a zároveň jsou opravovány chyby méně přesnějších modelů, tzv. boosting. Tyto různorodé modely se vytváří změnou trénovacích množin dat nebo měněním parametrů modelu. [21]

# 4 Metody detekce drift konceptu

V první řadě bakalářská práce pojednává o metodách zabývající se statistickou regulací procesů operující s chybovostí detekce vztahující se ke směrodatné odchylce. Hoeffdingova metoda vyniká tím, že je nezávislá na pravděpodobnostním rozložení dat a lze tak pozorovat odchylky některých vzorků od pravděpodobnostní distribuce. Poslední zde zmíněnou metodou je přístup kumulativního součtu, který patří mezi nástroje sekvenční analýzy. Na obrázku 4.1 jsou výše zmíněné metody detekce drift konceptu schématicky znázorněny.




Obr. 4.1: Schéma rozdělení metod pro detekci drift konceptu

## 4.1 Statistická regulace procesů

Statistická regulace procesů operuje s chybovostí detekce vztahující se ke směrodatné odchylce. Patří sem metoda detekce driftu, včasná detekce driftu a detekce využívající statistického testování. Další z uvedených metod je vážená statistická regulace procesů, která navíc míry chybovosti penalizuje vážným průměrem.

### 4.1.1 Metodu detekce driftu (DDM)

DDM je jednou z prvních používaných metod detekce drift konceptu. Její princip spočívá ve sledování míry chybovosti klasifikace $p_i$, kterou chápeme jako pravděpodobnost nesprávné predikce a její směrodatné odchylky $\sigma_i$, kterou definujeme jako [3]:

$$\sigma_i = \sqrt{\frac{p_i\,(1-p_i)}{i}}, \tag{4.1}$$

kde pro směrodatnou odchylku $\sigma_i$ platí, že s rostoucí posloupností výstupních dat $i$ se bude snižovat míra chybovosti $p_i$ [3] [6]. Ke sledování drift konceptu slouží dvě proměnné $p_{min}$ a $\sigma_{min}$, které se aktualizují v případě, že došlo ke změně v konceptu. Studie od Pesaranghadera A. a spol. [22] uvádí tři možné případy:

1. Drift koncept ještě nenastal, pak:

$$p_i + \sigma_i < p_{\min} + \sigma_{\min}. \tag{4.2}$$

2. Dalším případem je, že došlo k překročení varovné úrovně:

$$p_i + \sigma_i \geq p_{\min} + 2 \cdot \sigma_{\min} \tag{4.3}$$

3. Změnu distribuce dat v metodě detekce driftu detekujeme, když:

$$p_i + \sigma_i \geq p_{\min} + 3 \cdot \sigma_{\min}. \tag{4.4}$$

### 4.1.2 Metoda včasné detekce driftu (EDDM)

EDDM analyzuje vzdálenosti mezi dvěma po sobě následujícími chybami klasifikace. Nevýhodou této metody je časová prodleva, kdy se nejprve musí detekovat minimálně 30 chyb klasifikace [3]. Až poté dochází k případné signalizaci změny v konceptu, což by mohl být problém hlavně pro nevyvážená data. Na druhou stranu je metoda EDDM výhodnější pro pozvolné změny, kdy při dostatečném nasycení výstupních dat dochází ke spolehlivému rozložení chyb klasifikace. Tedy detekci postupných drift konceptů na rozdíl od metody DDM, která je účinnější pro detekci náhlých drift konceptů. Pro EDDM platí [1] [3]:

1. K překročení varovné úrovně dojde, pokud:

$$\frac{(p'_i + 2 \cdot s'_i)}{(p'_{\max} + 2 \cdot s'_{\max})} < \alpha. \tag{4.5}$$

2. Ke změně distribuce dat dochází, když:

$$\frac{(p'_i + 2 \cdot s'_i)}{(p'_{\max} + 2 \cdot s'_{\max})} < \beta. \tag{4.6}$$

V rovnicích 4.5 a 4.6 $p'_i$ představuje průměrnou vzdálenost mezi chybami a zároveň $s'_i$ je jejich směrodatná odchylka. Když se v EDDM dosáhne maximální vzdálenosti mezi chybami a dvojnásobek maximální směrodatné odchylky, vytvoří se dvě nové proměnné $p'_{max}$ a $s'_{max}$, které se ukládají do paměti a slouží pro příští detekci drift konceptu. Poté pomocí vztahů 4.5 a 4.6 algoritmus EDDM vypočítá průměrnou vzdálenost, směrodatnou odchylku a výslednou hodnotu porovná s prahem $\alpha$. Pokud je hodnota vypočítaného výrazu dle nerovnice 4.5 menší než $\alpha$, dojde k upozornění. Když je hodnota vypočítaného výrazu uvedeného v nerovnici 4.6 menší než $\beta$, dojde ke změně distribuce dat. Následně se model znovu přeučí a tím pádem se změní hodnoty proměnné $p'_i$ a $s'_i$. [1] [3]

### 4.1.3 Statistické testování

Detekce využívající statistického testování za předpokladu stejných proporcí a korekcí kontinuity pomocí dvou odlišných oken. Tato metoda sleduje poměr nedávné ku celkové přesnosti. Platí zde vztah:

$$\mathrm{T}\left(r_0, r_r, n_0, n_r\right) = \frac{\left|\frac{r_0}{n_0} - \frac{r_r}{n_r}\right| - \frac{1}{2}\left(\frac{1}{n_0} + \frac{1}{n_r}\right)}{\sqrt{p'\left(1-p'\right) \cdot \left(\frac{1}{n_0} + \frac{1}{n_r}\right)}}, \tag{4.7}$$

kde $r_0$ je celkový počet správných klasifikací kromě příkladů $W$, kde $r_r$ představuje počet správných klasifikací v případě, že $W = n_r$ a $p'$ se rovná $\frac{r_0+r_r}{n_0+n_r}$. Výsledná hodnota se porovná s percentilem normálního rozložení, abychom získali p-hodnotu. Když je zjištěno, že p-hodnota je nižší než $T$, potom $\frac{r_0}{n_0} < \frac{r_r}{n_r}$. Z čehož vyplývá, že celková přesnost je menší než přesnost předešlá, v důsledku toho je detekován drift koncept. [23]

### 4.1.4 Vážená statistická regulace procesů

Jeden ze starších přístupů je detekce drift konceptu pomocí exponenciálně váženého klouzavého průměru. Tato metoda oproti výše zmíněným ještě navíc míry chybavosti penalizuje váženým průměrem, který umožňuje rychlejší aktualizaci odhadu chyb. Starším pozorováním je přidělena nižší váha klesající exponenciálně. Nicméně, tento přístup se jeví efektivní pouze za určitých předpokladů. Metoda vážené statistické regulace procesů předpokládá znalost míry chybovosti, která v praxi většinou chybí. Předpokládá se, že míra chybovosti se směrodatnou odchylkou budou odhadovány v reálném čase. [24]

Zavedeme binární parametr $p_t$, který představuje pravděpodobnost nesprávné klasifikace objektu v čase $t$. Tento parametr může nabývat pouze dvou hodnot $p_0$ nebo $p_1$, což jsou pravděpodobnosti před a po detekci drift konceptu. Na základě těchto znalostí lze odhad rozptylu $Z_t$ popsat jako:

$$\sigma^2_{Z_t} = \sqrt{p_0\left(1-p_0\right)\frac{\lambda}{\lambda-1}\left(1-(1-\lambda)^{2t}\right)}, \tag{4.8}$$

kde $\lambda$ je koeficient zapomínání. Nicméně, pokud neznáme pravděpodobnost chybovosti před driftem $p_0$, zavádíme nový odhad:

$$p_{0,t} = \frac{1}{t}\sum_{i=1}^{t} X_i = \frac{t-1}{t} p_{0,t-1} + \frac{1}{t} X_t. \tag{4.9}$$

Drift koncept nastane právě tehdy, když odhad $Z_t$ překročí prahovou hodnotu, která je dána nerovnicí:

$$Z_t > p_{0,t} + L\sigma_{Z_t}, \tag{4.10}$$

kde $L$ je kontrolní parametr, který říká, jak dalece se liší odhad $Z_t$ od společného průměru náhodných proměnných před změnou konceptu. Poslední neznámou veličinou je směrodatná odchylka $\sigma_{Z_t}$:

$$\sigma_{Z_t} = \sqrt{\frac{\lambda}{2-\lambda}\left(1-(1-\lambda)^{2t}\right)}\,\sigma_x, \tag{4.11}$$

kde $\sigma_x$ je známá směrodatná odchylka. [25]

## 4.2 Hoeffdingova metoda

Metoda založená na Hoeffdingově teorému nerovnosti říká, že pro $n$ náhodných nezávislých vstupních vzorků $X_n$, které s určitou pravděpodobností $\delta$ nabývají hodnot predikce 0, nebo 1. Potom rozdíl průměrů přes všechny empirické $n$ a skutečným průměrem bude větší než hodnota Hoeffdingovo číslo $\varepsilon_H$. Tato metoda vyniká vlastností, že Hoeffdingova nerovnost je nezávislá na pravděpodobnostním rozdělení dat a díky tomu každému vzorku $n$ přiřadí k horní hranici odchylky střední hodnotu. Na základě výpočetní náročnosti se Hoeffdingova metoda řadí mezi rychlou detekci drift konceptu. Hoeffdinovo číslo $\varepsilon_H$ se vypočte pomocí Hoeffdingova teorému [22]:

$$\varepsilon_H = \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}} \tag{4.12}$$

Algoritmus založený na Hoeffding metodě rychlé detekce drift konceptu posune klouzavé okno o dané velikosti po objektech predikce. Dle nerovnosti pošle na výstup okna 1, pokud je výsledek predikce pravdivý, v opačném případě 0. Poté porovná pravděpodobnost výskytu 1 v klouzavém okně za čas $t$ ($p_t^1$) s maximální pravděpodobností výskytu 1 ($p_{\max}^1$). Pokud je hodnota $p_t^1$ větší než $p_{\max}^1$ dojde k posunu okna a hodnota $p_{\max}^1$ bude přepsána. Předpokladem je, aby se přesnost predikce s nárustem počtu vzorků zvyšovala, nebo zůstala stabilní. Při poklesu $p_t^1$ a zvyšující se $p_{\max}^1$ roste pravděpodobnost, že by mohlo dojít ke změně konceptu. Při výrazném rozdílu $p_{\max}^1$ a $p_t^1$, kdy je hodnota rozdílu větší než Hoeffdingovo číslo $\varepsilon_H$, které reprezentuje práh, dochází definitivně k lokalizaci drift konceptu. [22]

## 4.3 Přístupy kumulativního součtu

Detekce drift konceptu na základě kumulativního součtu patří mezi nástroje sekvenční analýzy. To znamená, že dochází k postupnému vyhodnocování výsledků predikce tak, jak sekvenčně přichází za sebou.

Samotná metoda monitoruje průměrný odhad $\bar{x}_t$ z pozorovaných dat $x_t$, kde čas $t \in (-\infty, 0)$. Mínus nekonečno označuje, že se jedná o dlouhodobé pozorování

historických dat a v čase 0 jde o odhad z aktuálního časového okamžiku. Poté je dle vzorce vypočten kumulativní součet $C_t$ [26]:

$$C_t = \sum_{t=-\infty}^{0} (x_t - \bar{x}_t) \tag{4.13}$$

Odečtením průměrné hodnoty odhadu dostaneme lokální variaci dat bez rušivých globálních variací, což umožňuje tyto variace dále analyzovat. Pokud kumulativní součet kolísá kolem nuly, jedná se o statická data se stabilním charakterem. V případě, kdy má tento součet pozitivní trend $\overline{x_{t+1}} > \bar{x}_t$ nebo negativní $\overline{x_{t+1}} < \bar{x}_t$, musí být zavedena prahová hodnota pro stanovení horní hranice $\lambda_P$ a stejně tak pro dolní hranici $\lambda_N$ definující přijatelnou míru variace dat. Když dojde k překročení prahové hodnoty pro $\lambda_P$ nebo $\lambda_N$, signalizuje se změna distribuce dat a detekuje se koncept drift. [26]

## 4.4 Ostatní principy detekce

Mezi ostatní principy detekce drift konceptu byla zařazena metoda detekce driftu v reálném čase, která signalizuje změnu v distribuci dat na základě shlukovací analýzy. Další je metoda A-vzdálenosti detekující drift koncept přes diferenci mezi dvěma okny. Závěrem je uvedena metoda hustoty okrajů, jenž odhaluje drift koncept pomocí podpůrných vektorů.

### 4.4.1 Detekce driftu v reálném čase

Metoda detekce drift konceptu v reálném čase bývá používána v případech, kdy nejsou známa cílová data. Například, kvůli velkému rozsahu vstupních dat, nelze jako u předchozích metod spoléhat na predikční modely. Tyto algoritmy fungují na principu shlukování pomocí metody k-means, kde $k$ udává počet shluků, které jsou reprezentovány svými těžišti, tzv. centroidy. Poté jsou vypočítávány vzdálenosti mezi centroidy a objekty. Následně jsou vypočítané vzdálenosti mezi sebou porovnány a objekty jsou přiděleny k nejbližším centroidům. [27]

Nejvzdálenější objekty od centroidů pak tvoří hranici v rámci daného shluku. Když se naskytne objekt ležící v těsné blízkosti hranice shluku, označuje se jako inovativní. Tento objekt nemusí nutně znamenat, že došlo k drift konceptu, protože se může klidně jednat i o nově vznikající shluk. Proto se tyto nejednoznačné objekty ukládají do krátkodobé paměti a podléhají monitorování, zda nevznikají nové shluky, či se objekt k některému nepřidal. Pokud i v dalších iteracích je objekt bez shluku, lze usuzovat, že tento objekt nemá společné příznaky s ostatními objekty a považujeme ho za koncept drift. Probíhá pravidelná aktualizace modelu za účelem

udržení normálního konceptu. Aktualizace probíhá tehdy, když v krátkodobé paměti je uložen konkrétní počet nejednoznačných objektů, který nastaví uživatel. Na základě toho dokáže algoritmus v reálném čase začlenit validní objekty do shluků, což je žádoucí. [27]

### 4.4.2 A-vzdálenost

Metodou založenou na A-vzdálenosti je detekována změna distribuce dat pomocí vzdáleností rozdílů mezi dvěma okny představující pravděpodobnostní rozdělení náhodné proměnné, které do těchto oken náleží. A-vzdálenost je tedy vzdáleností mezi těmito dvěma okny, která signalizuje, k jak moc velké diferenci v distribuci dat došlo. [28]

Jedno z oken je stacionární a druhé se posouvá po vzorcích směrem doprava, tudíž je dynamické. Obě okna jsou dále rozdělena na intervaly. Výsledek diference se porovnává se stanovenými intervaly, jestliže je náhodná hodnota rozdílu menší, spadá do tohoto intervalu. Pokud je hodnota diference větší než náležitý interval, dochází k signalizaci drift konceptu. [28]

### 4.4.3 Metoda hustoty okrajů

Metoda hustoty okrajů je založena na podpůrných vektorech. Tato metoda bývá implementována právě tehdy, když se jedná o pravděpodobnostní klasifikátor, který generuje odhady pravděpodobnosti rozdělení objektů do tříd. [29]

Lze si tedy představit pásmo mezi dvěma třídami, tzv. nadrovinu. Nadrovina umožňuje určit chybu klasifikace vstupních objektů $x_i$ a popřípadě převést objekty se zápornými souřadnicemi na kladné souřadnice pomocí operátoru $\xi_i$. Operace převedení záporných souřadnice na kladné je umožněna jen tehdy, pokud se jedná o tzv. měkké okraje, které tohle umožnují. Metoda hustoty okrajů je dána svou optimalizační funkcí:

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i. \tag{4.14}$$

Zároveň pro rovnici 4.14 platí dvě podmínky:

$$y_i \left( x_i^T w + b \right) \geq 1 - \xi_i, \tag{4.15}$$

$$\xi_i \geq 0, \tag{4.16}$$

kde $y_i$ označuje výstup, $C$ je konstanta vyjadřující váhu chybné klasifikace, $w$ je normálový vektor určující váhu vstupního objektu $x_i^T$ a $b$ je práh. [29]

# 5 Metody korekce drift konceptu

Algoritmy korekce slouží k úpravě detekovaných koncepčních driftů na základě ověření správnosti detekce. První z algoritmů je založen na metodě ADWIN a dokáže ihned zareagovat na změnu v rozložení dat. Zatímco metoda korekce přes vektorovou regresi provádí korekci v pravidelných intervalech.

## 5.1 ADWIN

Metoda ADWIN je založená na adaptivním posuvném okně. Tato sofistikovaná metoda využívá přizpůsobivého klouzavého okna proměnlivé velikosti, která se adaptuje podle rychlosti změny rozložení dat v náležitém okně. Algoritmus založený na tomto principu je schopen odhalit i falešně pozitivní či falešně negativní drift koncepty. [30]

Pokud není změna v konceptu detekována, tak se okno dynamicky zvětšuje a zároveň se i zbavuje starých dat, takže logaritmicky klesá výpočetní náročnost. V opačném případě, když je změna rozložení dat zaznamenána, okno se zmenšuje. Algoritmus přidá $n$-tici dat do adaptivního okna. Následuje analýza $n$ vzorků za účelem detekce drift konceptu. [30]

K zpřesnění výpočtu jsou zavedena dvě podokna *W1* a *W2*, která se určují přes všechny možné kombinace podoken, aby bylo dosaženo co možno nejpřesnějších výsledků. Poté se porovnají distribuce dat v obou podoknech a pokud jsou průměrné hodnoty dat odlišné, dochází ke korekci drift konceptu. Nejstarší $n$-tice se odstraní a dochází k rozšiřování okna. Okno se zvětšuje a odstraňuje staré $n$-tice do té doby, než nedojde k další korekci drift konceptu. [31]

## 5.2 Vektorová regrese

Algoritmy založené na podpoře vektorové regrese patří k přístupům učení s učitelem, které na základě symetrického odhadu penalizují chybné odhady. Touto metodou lze dosáhnout vysoké přesnosti predikce na základě zobecnění klasifikačního problému odhadem spojité vícerozměrné funkce.

Podpora vektorové regrese reprezentuje tedy optimalizační problém, ve kterém je snaha nalézt místo nadroviny, jež je v daném místě nejužší se středem na povrchu nadroviny. Tím se zároveň minimalizuje chyba predikce. Aproximační dvourozměrná funkce $f(x, w)$ je definována vztahem:

$$f(x, w) = \sum_{\mathrm{i}=1}^{M} w_i x^i, \tag{5.1}$$

kde $M$ je řád polynomu, $x$ jsou vstupní data a hodnota $w$ udává šířku nadroviny. [32]

S rostoucím řádem polynomu klesá chyba, ale zároveň lze očekávat vysokou výpočetní náročnost. Za optimální se nepožadují v některých případech ani nízké řády. Například polynom prvního řádu, tedy lineární funkce, nemusí dosahovat očekávaných výstupních hodnot. Na základě toho je nutno počítat s vyšší chybovostí predikce. Polynom šestého řádu se uvádí jako optimální a je brán jako kompromis mezi ideální šířkou nadroviny a mírou chybné predikce. Dalším parametrem je ztrátová funkce $\epsilon$. Tento parametr udává rozdíl mezi očekávaným a reálným výstupem. Pokud je rozdíl menší než $\epsilon$, dochází ke snížení citlivosti pro detekci chyby a tím pádem je kolem nadroviny více objektů, tedy podpůrných vektorů. Naopak při zvýšení diference bude algoritmus citlivější a je daleko větší pravděpodobnost, že bude detekován koncept drift. [32]

# 6 Dataset

Pro analýzu drift konceptu byly použity dva různé datasety. První dataset reprezentuje uměle vytvořený fantomový drift koncept. Tento dataset byl použit pro testování detekčních a korekčních algoritmů drift konceptu. Na základě tohoto testování byly vybrány vhodné detekční metody za účelem detekce drift konceptu v metabolomické analýze. Druhý dataset již obsahoval reálná data, kterými byly koncentrace metabolitů.

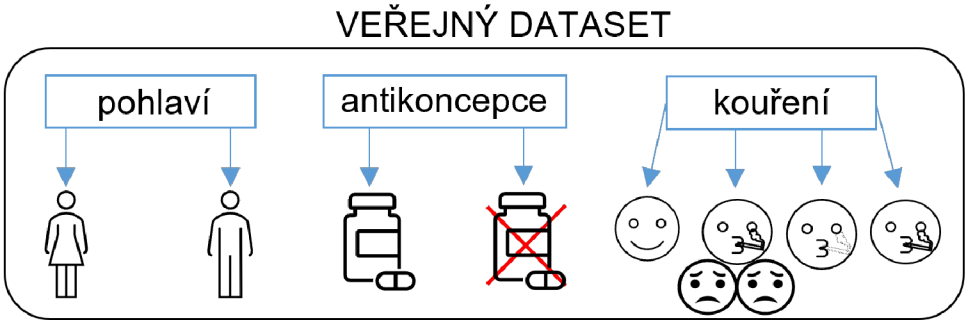
## 6.1 Fantomová data

Pro testování algoritmů na analýzu drift konceptu byl vybrán spamový dataset [33]. Tento dataset obsahuje 6213 vstupních objektů a 499 příznaků.

Dataset byl rozdělen na vstupní a cílová tzv. očekávaná data. Očekávaná data reprezentují dvě třídy - 0, jestliže se nejedná o spam, nebo 1, pakliže se jedná o spam. Vstupní data v datasetu jsou taktéž binární.

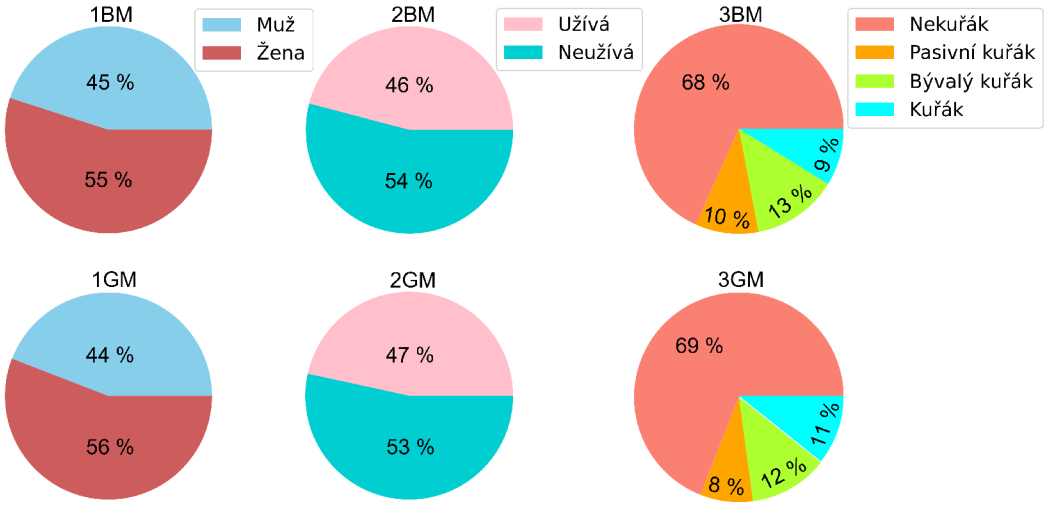
## 6.2 Reálná data

Pro analýzu drift konceptu v metabolomické analýze byly použity datasety pocházející z kohortové studie od Chu a spol. [34]. Získaná data pocházejí od 534 objektů rodu *Homo sapiens*. Konkrétně byly měřeny tělní metabolity od 237 mužů a 296 žen ve věku od 18 do 75 let. Studie přinesla dva metabolomické datasety lišící se použitými technikami měření. První dataset nese označení BM (z angl. *Brainshake Metabolomics*). Data pro BM byla měřena technikou spektroskopie nukleární magnetická rezonance, jenž je blíže popsána v kapitole 2.1 Výsledky měření přinesly 231 příznaků s 200 absolutními koncentracemi. Další dataset nesoucí název GM (z angl. *General Metabolomics*) zahrnuje 1589 příznaků s 257 absolutními koncentracemi, které byly měřeny technikou hmotnostní spektrometrie doby letu. O hmotnostní spektrometrii doby letu pojednává kapitola 2.1. Poslední, avšak nezbytný, veřejný dataset obsahuje osobní iniciály, a hlavně očekávaná data pro jednotlivé objekty. Těmito očekávanými daty jsou tři různé klasifikační problémy, jak můžeme vidět na obrázku 6.1. První klasifikační problém označuje pohlaví (žena/muž). Další klasifikační problém ukazuje, zda žena užívá antikoncepci (užívá/neužívá). Poslední klasifikační problém obsahuje čtyři třídy (nekuřák/pasivní kuřák/bývalý kuřák/aktivní kuřák).




Obr. 6.1: Schéma znázorňující tři klasifikační problémy z datasetu dostupného [34]

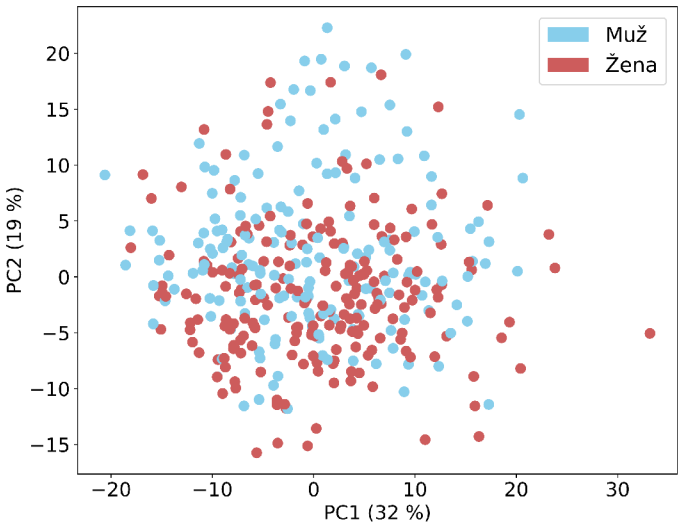
Zároveň byly vytvořeny grafy vyjadřující četnosti jednotlivých tříd pro každý klasifikační problém v procentech. Jak můžeme vidět na obrázku 6.2 – očekávaná data pro pohlaví a užívání antikoncepce jsou relativně vyvážená. To se ovšem nedá říct o očekávaných datech pro klasifikaci kuřáků, kdy můžeme zaznamenat nevyváženost, co se týče četnosti jednotlivých klasifikačních tříd.



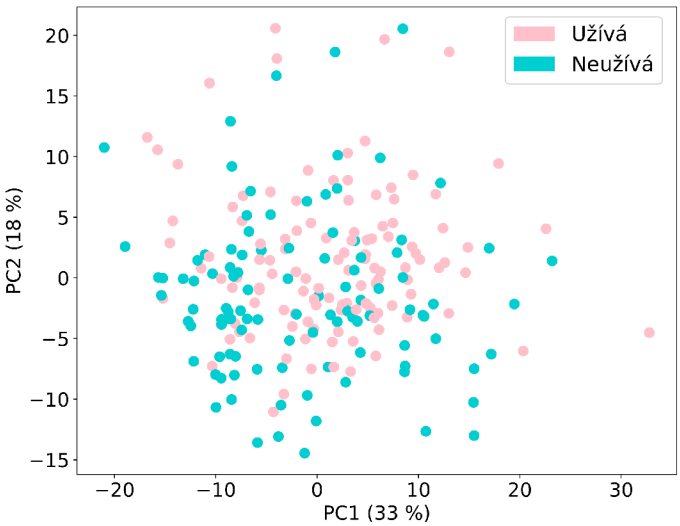

Obr. 6.2: Grafy vyjadřující četnosti tříd pro daný klasifikační problém a dataset.
Klasifikčaní problém: 1 – pohlaví, 2 – užívání antikoncepce, 3 – kouření.
Dataset: BM - první tři horní grafy, GM - spodní tři grafy

Oba datasety byly standardizovány, čímž bylo dosaženo nezávislosti příznaků na různorodých jednotkách a zároveň bylo sjednoceno měřítko [35].Vzhledem k vysokodimenzionalitě vstupních dat, nelze datasety jednoduše vizualizovat. Jako jedním z alternativních řešení redukce dimenzionality se jevila analýza hlavních komponent (PCA) sloužící nejen pro redukci demenzionality vstupních dat, ale i pro vizualizaci vstupních dat. Vstupní data byla pomocí PCA redukována do dvou dimenzí a vykreslena [36]. Na obrázku 6.3 vidíme graficky znázorněnou výslednou analýzu

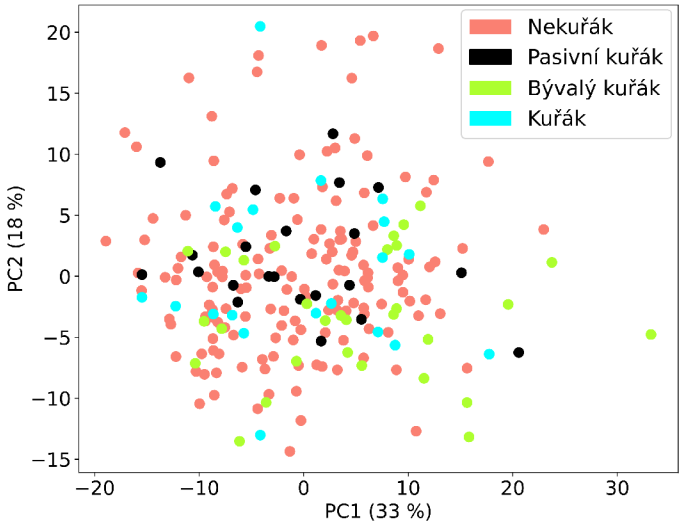
hlavních komponent pro dataset BM s vyznačenými očekávanými daty pro klasifikaci pohlaví. Obdobně je tomu na obrázku 6.4 pro klasifikaci užívání antikoncepce. Posledním klasifikačním problémem bylo kouření, kdy na obrázku 6.5 jsou vyznačeny čtyři výstupní třídy. PCA prezentující výše zmíněné klasifikační problémy pro GM dataset jsou umístěny v příloze A.




Obr. 6.3: Vizualizace analýzy hlavních komponent BM datasetu s vyznačenými očekávanými daty pro klasifikaci pohlaví




Obr. 6.4: Vizualizace analýzy hlavních komponent BM datasetu s vyznačenými očekávanými daty pro klasifikaci užívání antikoncepce

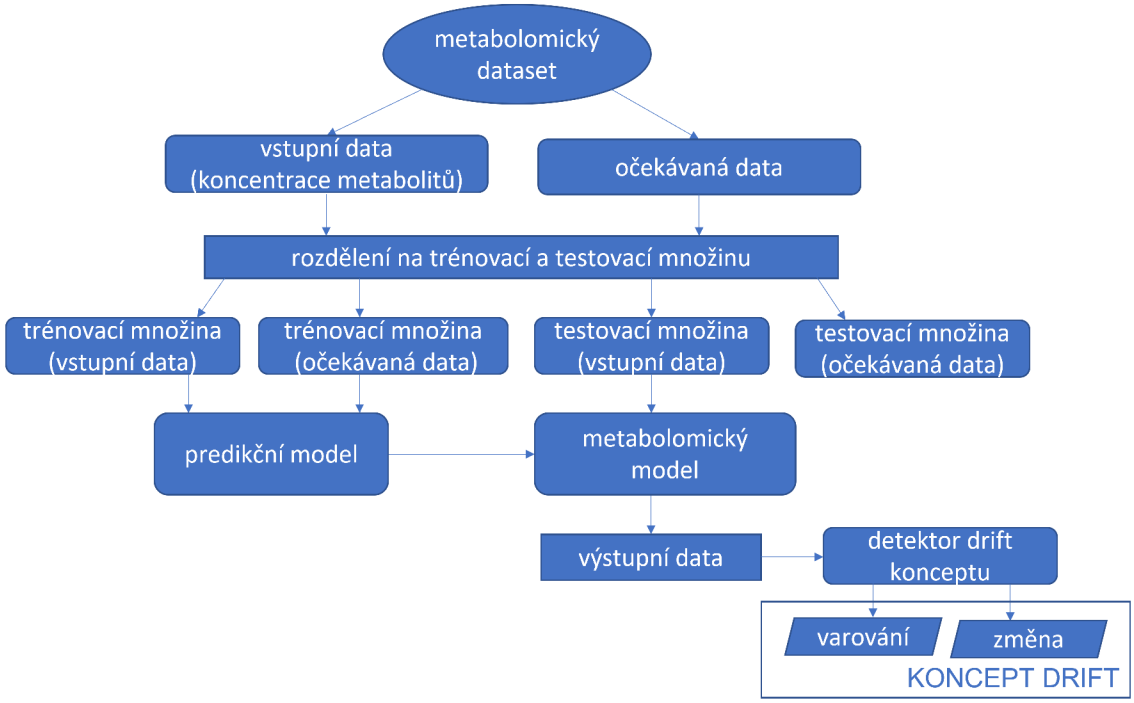



Obr. 6.5: Vizualizace analýzy hlavních komponent BM datasetu s vyznačenými očekávanými daty pro klasifikaci kuřáků

Analýza hlavních komponent ukázala, že se nevytvořly viditelně ohraničené shluky objektů. Což může být způsobeno vzájemnými vnitřními interakcemi mezi jednotlivími metabolity. Vlivem výběru vstupních dat pro klasifikaci pohlaví statistickým testem analýza rozptylu (ANOVA) [37], došlo ke ztrátě některých původně detekovaných drift konceptů a přesnost modelů se nezměnila. Proto byla data ponechána v původní neredukované podobě. Na druhou stranu, vstupní data, jež byla použita pro klasifikaci užívání antikoncepce a kouření, byla z důvodu nizké přesnosti klasifikace redukována. Pro selekci vstupních dat byly otestovány dva přístupy. První z nich je již zmiňovaný statistický test ANOVA pro experimentálně definovaný počet příznaků. Z BM datasetu bylo vybráno pro predikci užívání antikoncepce 25 příznaků a 50 příznaků pro predikci kouřáků. Obdobně tomu bylo u GM datasetu, kdy bylo vybráno 100 příznaků, jak pro predikci užívání antikoncepce, tak i pro predikci kuřáků. Druhý přístup spočíval v korelaci vstupních dat [38]. Pro každou možnou dvojici příznaků byla vypočtena hodnota Pearsonova korelačního koeficientu. Pokud hodnota Pearsonova korelačního koeficientu byla větší než zadaný práh, příznaky s takovou hodnotou byly považovány za redundantní a byly vyřazeny ze vstupních dat. Ani přes testování korelace příznaků pro prahové honoty 0.6, 0.7, 0.8, 0.9, a jejich následné redukce, nedošlo ke zvýšení přesnosti modelů. Z důvodu výše popsaných skutečností, byl pro selekci vstupních dat vybrán statistický test ANOVA, čímž došlo ke zvýšení přesnosti modelů a počet detekovaných drift konceptů se nelišil od původních neredukovaných dat.

# 7 Implementace detektorů drift konceptu

Detekce drift konceptu na predikčních modelech byla rozdělena na dvě části. První část je testovací a pojednává o detekci drift konceptu v predikcích na fantomových datech. Druhá část je věnována detekci drift konceptu v metabolomických predikčních modelech. Princip detekce drift konceptu v predikcích metabolomických modelů je znázorněn na obrázku 7.1 popisující námi zvolený postupu řešení.




Obr. 7.1: Schéma zvoleného postupu řešení

Analýza drift konceptu byla realizována ve skriptovacím jazyce Python verze 3.7 [39]. Tento programovací jazyk byl vybrán pro jeho jednoduchost, všestrannost a popularitu. Vývojovým prostředím byl zvolen Google Colaboratory, který vyniká svou buňkovou strukturou, tudíž je přehledný a praktický. Pro analýzu dat byly zvoleny následující knihovny:

- Pandas [40] – pro manipulaci a úpravu dat,
- Numpy [41] – nabízí řadu užitečných matematických operací,
- Matplotlib [42] – vytváření grafů,
- Sklearen [43] – modelování na základě algoritmů strojového učení,
- Skmultiflow [44] – pro analýzu drift konceptu.

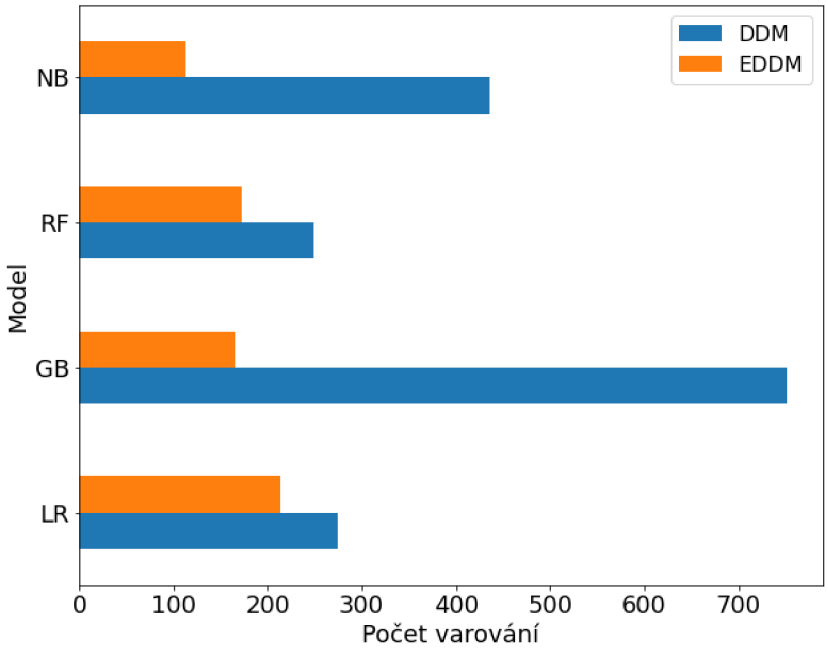
## 7.1 Detekce drift konceptu na fantomových datech

Pro testování algoritmů na detekci a korekci drift konceptu byly natrénovány modely na spamovém datasetu [33]. Spamový dataset byl zvolen na základě dynamického charakteru dat s cílem detekce změn v distribuci dat v predikcích natrénovaných modelů.

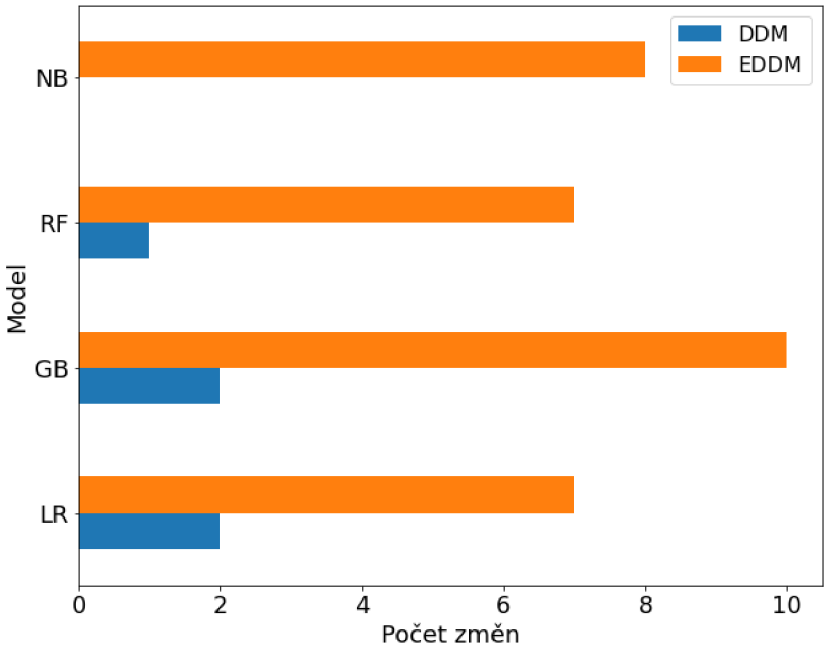
Dataset byl rozdělen na trénovací a testovací množinu v poměru 4:1. Tedy 80 % datasetu tvořilo trénovací množinu, na kterých byly jednotlivé predikční modely natrénovány. A 20 % datasetu bylo použito k otestování již natrénovaných predikčních modelů.

Použité predikční modely jsou založeny na algoritmech: logistické regrese (LR) [45], gradient boostingu (GB) [46], náhodném lesu (RF) [47] a Naivním Bayesovském klasifikátoru (NB) [48]. Principy výše zmíněných metod jsou popsány v kapitole 3.

Pro detekci drift konceptu byly implementovány dvě metody ze Scikit-Multiflow – metoda detekce driftu (DDM) a metoda včasné detekce driftu (EDDM) [44]. Principy těchto metod byly uvedeny v kapitole 4.1 pojednávající o statistické regulaci procesů. Tyto dvě metody byly aplikovány na modelem vytvořené predikce s cílem detekce drift konceptu v predikcích jednotlivých predikčních modelů. Detekce drift konceptu byla rozdělena na dvě úrovně. První úroveň byla varovná a druhá úroveň signalizovala změnu v distribuci dat. Počet detekovaných varovných úrovní v predikcích modelů prezentuje graf na obrázku 7.2 a graf vykreslující počet detekovaných změn v distribuci dat je znázorněn na obrázku 7.3.




Obr. 7.2: Graf počtů detekovaných varovných úrovní v predikcích modelů




Obr. 7.3: Graf počtů detekovaných změn konceptu v predikcích modelů

V porovnání detekčních algoritmů drift konceptu na predikčních modelech plyne, že metoda EDDM je daleko citlivější na detekci změn v konceptu viz obrázek 7.3. Naopak metoda DDM je citlivější pro detekci varovných úrovní viz obrázek 7.2. Což může být způsobeno tím, že metoda EDDM má daleko ostřejší hranici mezi detekcí varovné úrovně a změny v distribuci dat. Proto je tato metoda vhodnější pro pozvolné změny v distribuci dat oproti metodě DDM, která je vhodnější pro náhlé změny.

Následně byl implementován ze Scikit-Multiflow korekční algortimus drift konceptů, jenž je založen na metodě adaptivního posuvného okna (ADWIN) modifikovaného modelem k-nejbližších sousedů (KNN) [44]. Metoda ADWIN byla popsána v kapitole 5.1. Nejdříve byla modelem KNN vytvořena predikce vstupních dat. Poté byla predikce námi navržených modelů (LR, GB, RF, NB) porovnána s predikcí modelu KNN. Výsledkem byla korekce detekovaných drift konceptů s výkonem 1, což značí, že byly opraveny všechny detekované drift koncepty s maximálním výkonem.

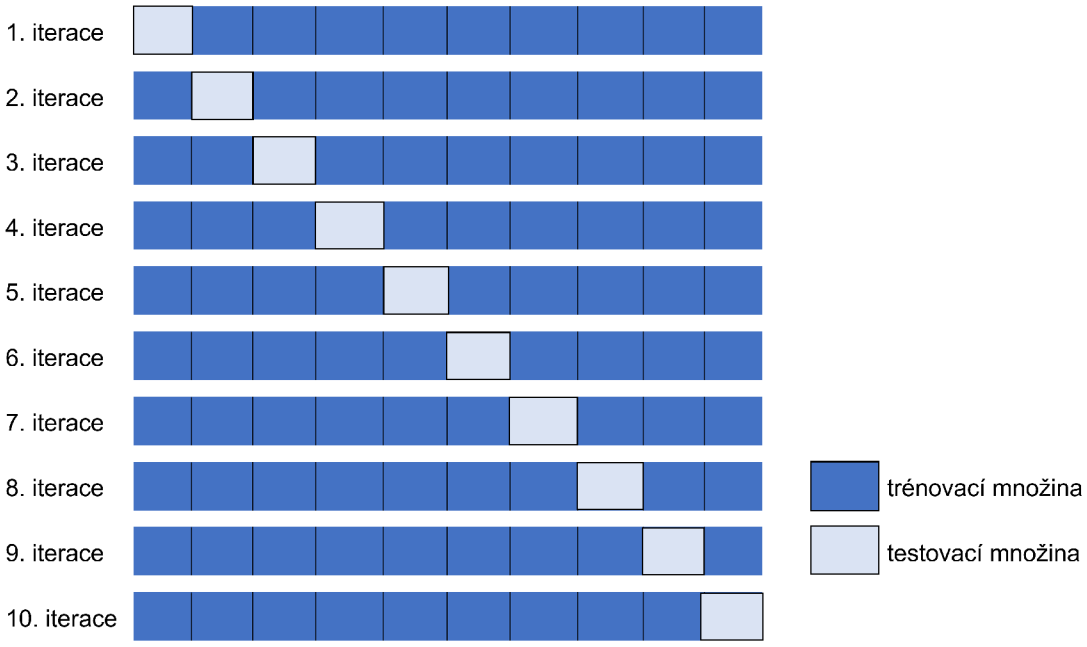
Na základě testování byly vybrány obě metody detekce drift konceptu s přihlédnutím k metodě EDDM, která detekovala více změn v konceptu než metoda DDM. Dalším důvodem preference detekčního algoritmu EDDM je jeho snížená citlivost pro detekci náhlých změn, které by mohly být příčinou chybovostí navržených predikčních modelů.

## 7.2 Detekce drift konceptu na metabolomických datech

Všechny zde zmíněné predikční modely byly natrénovány na koncentracích metabolitů v lidském těle pro rozpoznání pohlaví (žena/muž). Další skupina modelů klasifikovala, jestli žena užívá nebo neužívá antikoncepci. Poslední skupina modelů predikovala, zda dotyčná osoba kouří, kouřila, či pasivně inhaluje cigaretový kouř, nebo zda se jedná o bývalého kuřáka.

Použité predikční modely jsou založeny na algoritmech: logistické regrese (LR) [45], gradient boostingu (GB) [46], náhodném lesu (RF) [47] a Naivním Bayesovském klasifikátoru (NB) [48]. Principy výše zmíněných metod jsou popsány v kapitole 3.

U všech výše uvedených modelů byla pomocí desetinásobné křížové validace vyhodnocena průměrná přesnost predikce [49]. Což znamená, že trénovací a testovací množina dat byla rozdělena v poměru 9:1 a křížová validace se opakovala v deseti iteracích, jak můžeme vidět na obrázku 7.4. Zjištěná průměrná přesnost predikce pro každou datovou sadu BM a GM zvlášť je uvedena v tabulce 7.1 reprezentující tři klasifikační problémy.




Obr. 7.4: Desetinásobná křížová validace. Převzato a upraveno z [50]

Tab. 7.1: Průměrné hodnoty přesnosti jednotlivých metabolomických modelů predikujících pohlaví, užívání antikoncepce a kouření pro datasety BM a GM

| | **Pohlaví** | | | |
|---|---|---|---|---|
| | **Logistická regrese** | **Gradient boosting** | **Náhodný les** | **Naivní Bayes** |
| **BM** | 0.81 | 0.81 | 0.79 | 0.73 |
| **GM** | 0.83 | 0.90 | 0.87 | 0.76 |
| | **Antikoncepce** | | | |
| | **Logistická regrese** | **Gradient boosting** | **Náhodný les** | **Naivní Bayes** |
| **BM** | 0.84 | 0.80 | 0.77 | 0.75 |
| **GM** | 0.86 | 0.88 | 0.89 | 0.78 |
| | **Kouření** | | | |
| | **Logistická regrese** | **Gradient boosting** | **Náhodný les** | **Naivní Bayes** |
| **BM** | 0.70 | 0.64 | 0.68 | 0.42 |
| **GM** | 0.62 | 0.69 | 0.72 | 0.40 |

Další metrikou hodnocení přesnosti klasifikace bylo zvoleno F1-skóre porovnávající výstup predikce, který byl proveden na testovací množině vstupních dat, s testovací množinou dat očekávaných [51]. Výsledky F1-skóre jednotlivých modelů predikujících daný klsifikační problém pro dataset BM a GM zvlášť jsou uvedeny v tabulce 7.2.

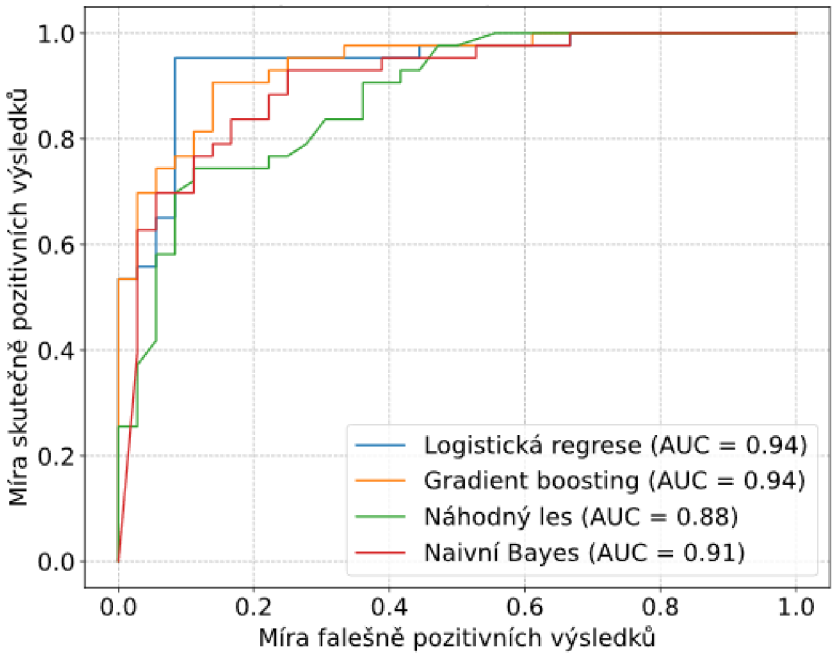
Tab. 7.2: Hodnoty F1-skóre jednotlivých metabolomických modelů predikujících pohlaví, užívání antikoncepce a kouření pro datasety BM a GM

| | **Pohlaví** | | | |
|---|---|---|---|---|
| | **Logistická regrese** | **Gradient boosting** | **Náhodný les** | **Naivní Bayes** |
| **BM** | 0.85 | 0.85 | 0.76 | 0.75 |
| **GM** | 0.89 | 0.87 | 0.88 | 0.73 |
| | **Antikoncepce** | | | |
| | **Logistická regrese** | **Gradient boosting** | **Náhodný les** | **Naivní Bayes** |
| **BM** | 0.87 | 0.76 | 0.75 | 0.75 |
| **GM** | 0.85 | 0.88 | 0.85 | 0.76 |
| | **Kouření** | | | |
| | **Logistická regrese** | **Gradient boosting** | **Náhodný les** | **Naivní Bayes** |
| **BM** | 0.70 | 0.67 | 0.70 | 0.59 |
| **GM** | 0.58 | 0.69 | 0.69 | 0.38 |

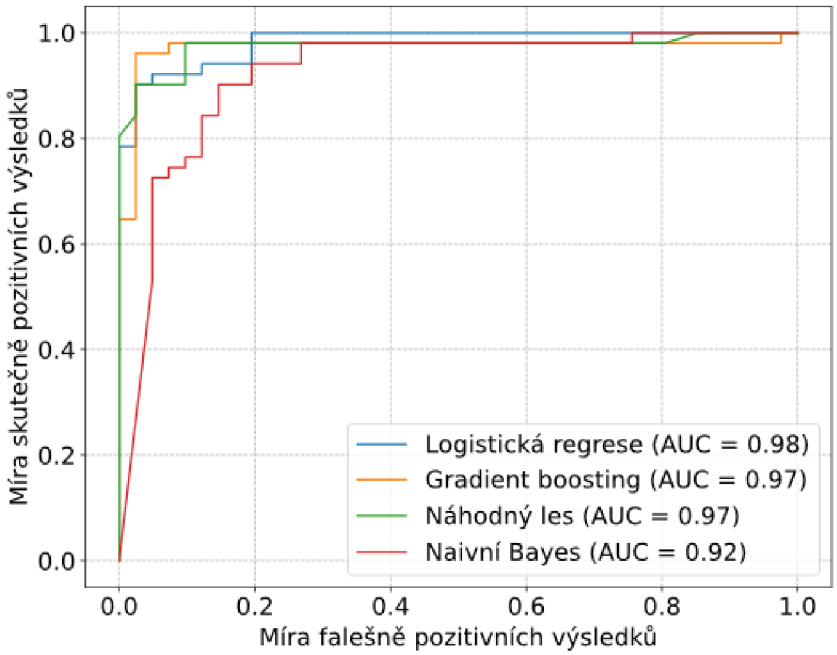
V případě predikce pohlaví, dle tabulky 7.1, dosáhl nejlepší průměrné přesnosti model založený na metodě gradient boosting v datové sadě GM. Tuto skutečnost

stvrzuje hodnota F1-skóre, která je v tabulce 7.2 pro tento případ taktéž jednou z nejvyšších. Dalším úkolem klasifikátorů byla predikce užívání antikoncepce, kde exceloval metabolomický model klasifikující na principu algoritmu náhodného lesa v datové sadě GM. V posledním případě byly modely natrénovány pro klasifikaci kuřáků do 4 tříd, kde uspokojivých výsledků dosáhl model založený na náhodném lesu. Důvodem tak nízkých hodnot přesností u klasifikace kuřáků byla značná nevyváženost očekávaných dat. V celkovém hodnocení nejvíce exceloval model založený na algoritmu gradient boosting, který lze považovat za nejsofistikovanější. Poté, co srovnáme obě tabulky, lze vyvodit, že hodnoty přesnosti a F1-skóre si téměř ve svém principu odpovídají. Tudíž, čím větší je přesnost predikce, tím větší je pravděpodobnost zařazení objektu do správné třídy.

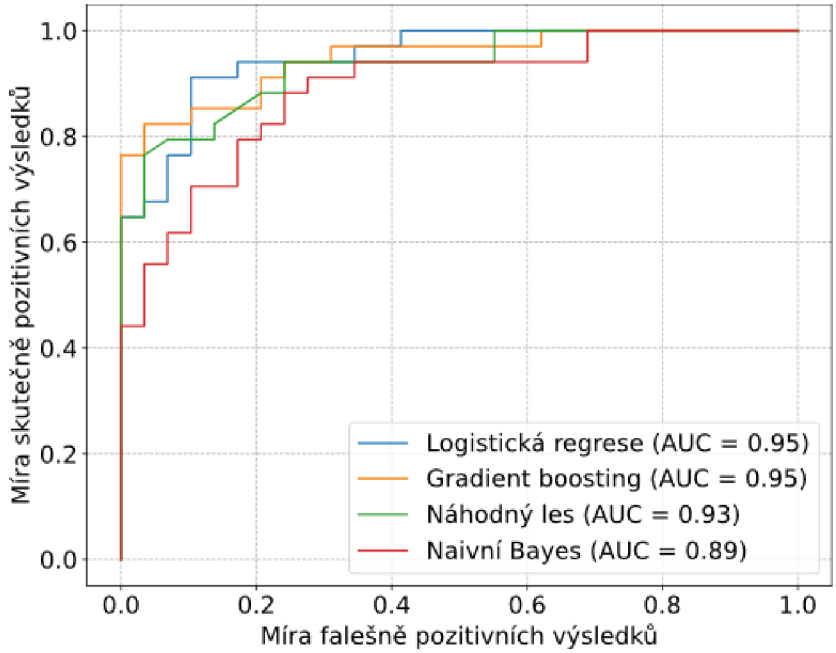
Poslední uvedenou metrikou pro hodnocení binární klasifikace modelů je ROC (z angl. *Receiver Operating Characteristic*) křivka. Jedná se o jednu z nejpoužívanějších metrik pro hodnocení kvality binární klasifikace. Udává vztah mezi senzitivitou, pro různé prahy klasifikace, a specificitou daného klasifikátoru. Jejím parametrem je plocha pod křivkou (AUC - z angl. *Area Under the Curve*), která objektivně hodnotí celý průběh křivky. Vykreslené průběhy ROC křivek s hodnotami AUC pro případ klasifikace pohlaví jsou uvedeny na obrázku 7.5 pro BM a na obrázku 7.6 pro GM dataset. Na obrázcích 7.7 pro BM a 7.8 pro GM, jsou průběhy ROC křivek pro případ klasifikace užívání antikoncepce. [52]



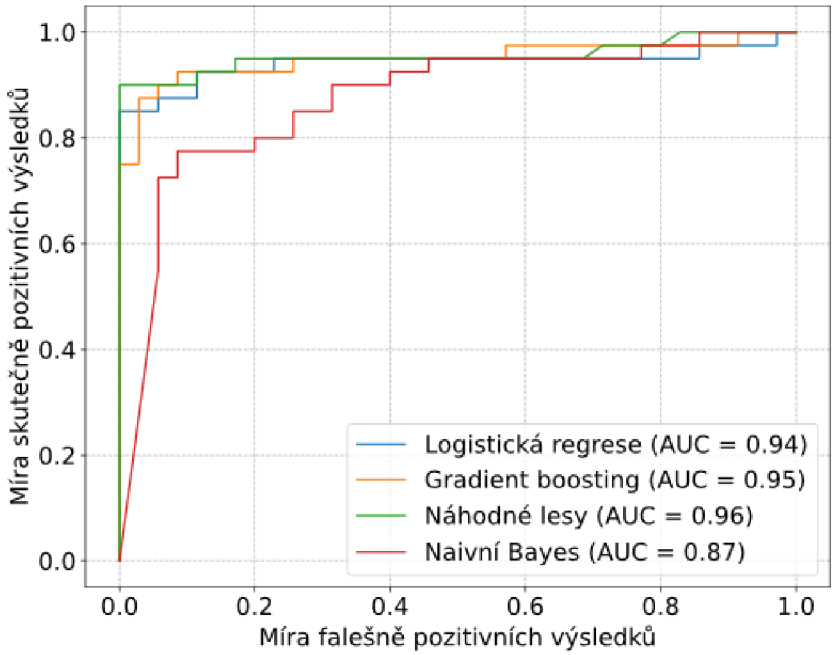

Obr. 7.5: Průběhy ROC křivek jednotlivých predikčních modelů klasifikujících pohlaví v BM datasetu a hodnoty AUC




Obr. 7.6: Průběhy ROC křivek jednotlivých predikčních modelů klasifikujících pohlaví v GM datasetu a hodnoty AUC




Obr. 7.7: Průběhy ROC křivek jednotlivých predikčních modelů klasifikujících užívání antikoncepce v BM datasetu a hodnoty AUC




Obr. 7.8: Průběhy ROC křivek jednotlivých predikčních modelů klasifikujících užívání antikoncepce v GM datasetu a hodnoty AUC

Na závěr byly porovnány jednotlivé hodnoty AUC ROC křivek napříč oběma datasety. Nejvyšší hodnoty AUC nabývaly predikční modely klasifikující v GM datasetu. Zatímco modely predikující v BM datasetu nabývaly nižších hodnot AUC. V praxi hodnoty AUC nabývají spojitých hodnot z rozsahu $< 0.5, 1 >$ a čím vyšší je hodnota AUC, tím větší byla úspěšnost klasifikace predikčního modelu.

Detekce drift konceptu v predikcích metabolomických modelů probíhala principiálně obdobně jako v případě testování detekčních algoritmů drift konceptu na fantomovém datasetu 7.1. Taktéž byla detekce drift konceptu rozdělena na dvě úrovně. První úroveň byla varovná a druhá úroveň signalizovala změnu v distribuci dat. Počet detekovaných varovných úrovní a změn v distribuci dat na daných metabolomických modelech metodou DDM a EDDM pro datovou sadu BM je uveden v tabulce 7.3 a pro dataset GM v tabulce 7.4.

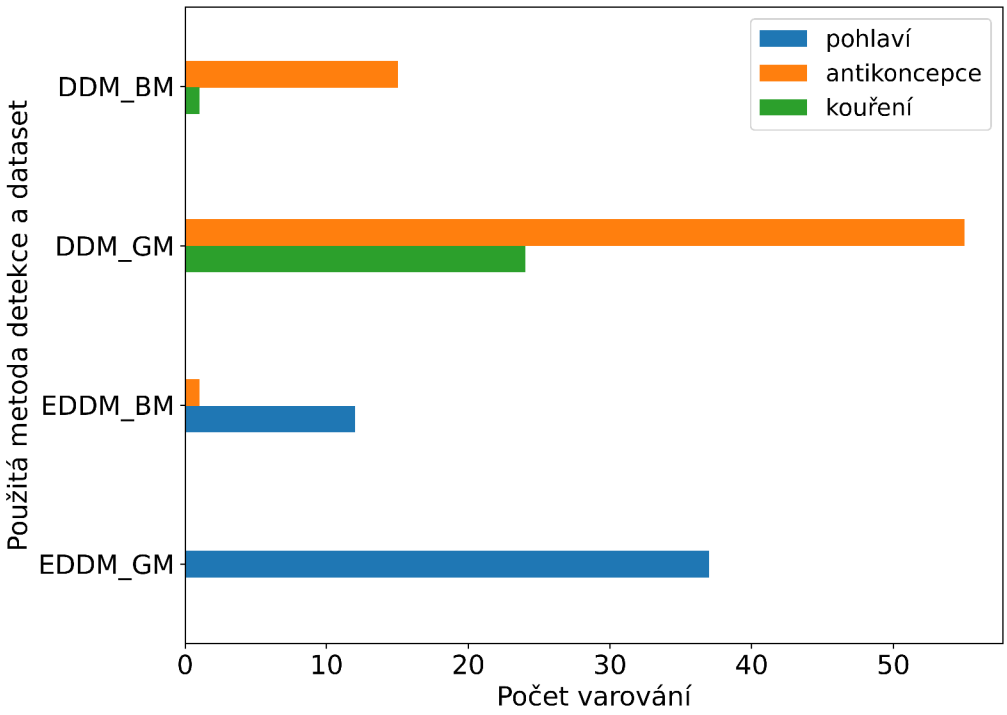
Tab. 7.3: Počet detekovaných varovných úrovní (V) a změn v distribuci dat (Z) na daných predikčních modelech metodou DDM a EEDM v BM datasetu (P-pohlaví, A-antikoncepce, K-kouření)

| | **DDM** | | | |
|---|---|---|---|---|
| | **Logistická regrese** | **Gradient boosting** | **Náhodný les** | **Naivní Bayes** |
| **P** | 0 V / 0 Z | 0 V / 0 Z | 0 V / 0 Z | 0 V / 0 Z |
| **A** | 0 V / 0 Z | 0 V / 0 Z | 15 V / 0 Z | 0 V / 0 Z |
| **K** | 0 V / 0 Z | 0 V / 0 Z | 0 V / 0 Z | 1 V / 0 Z |
| | **EDDM** | | | |
| | **Logistická regrese** | **Gradient boosting** | **Náhodný les** | **Naivní Bayes** |
| **P** | 0 V / 1 Z | 10 V / 1 Z | 2 V / 1 Z | 0 V / 1 Z |
| **A** | 1 V / 0 Z | 0 V / 0 Z | 0 V / 0 Z | 0 V / 0 Z |
| **K** | 0 V / 0 Z | 0 V / 0 Z | 0 V / 0 Z | 0 V / 0 Z |

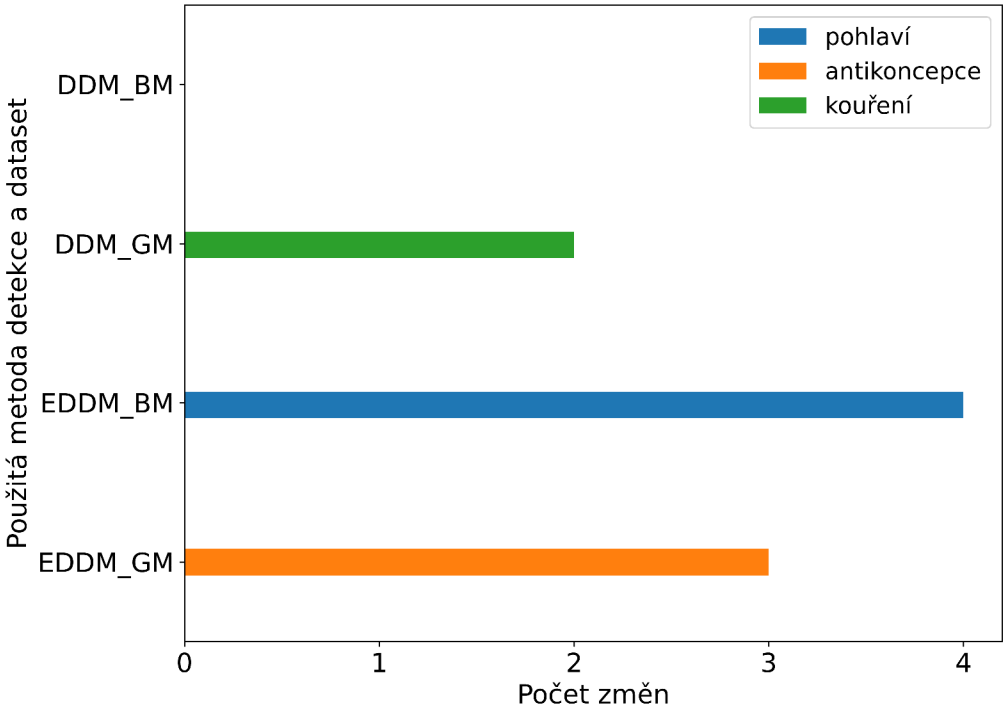
Tab. 7.4: Počet detekovaných varovných úrovní (V) a změn v distribuci dat (Z) na daných predikčních modelech metodou DDM a EEDM v GM datasetu (P-pohlaví, A-antikoncepce, K-kouření)

| | **DDM** | | | |
|---|---|---|---|---|
| | **Logistická regrese** | **Gradient boosting** | **Náhodný les** | **Naivní Bayes** |
| **P** | 0 V / 0 Z | 0 V / 0 Z | 0 V / 0 Z | 0 V / 0 Z |
| **A** | 0 V / 0 Z | 19 V / 0 Z | 15 V / 0 Z | 21 V / 0 Z |
| **K** | 7 V / 1 Z | 3 V / 1 Z | 6 V / 0 Z | 8 V / 0 Z |
| | **EDDM** | | | |
| | **Logistická regrese** | **Gradient boosting** | **Náhodný les** | **Naivní Bayes** |
| **P** | 18 V / 0 Z | 11 V / 0 Z | 0 V / 0 Z | 8 V / 0 Z |
| **A** | 0 V / 1 Z | 0 V / 1 Z | 0 V / 1 Z | 0 V / 0 Z |
| **K** | 0 V / 0 Z | 0 V / 0 Z | 0 V / 0 Z | 0 V / 0 Z |

Pro srovnání celkového počtu detekovaných drift konceptů metodou DDM a EDDM v jednotlivých datasetech byly vytvořené sloupcové grafy. Graf na obrázku 7.9 vizualizuje celkový počet detekovaných varovných úrovní konkrétní detekční metodou v daném datasetu. Druhý graf na obrázku 7.10 prezentuje počet detekovaných změn v distribuci dat taktéž konkrétní detekční metodou v daném datasetu.




Obr. 7.9: Graf počtů detekovaných varovných úrovní pro danou detekční metodu a dataset



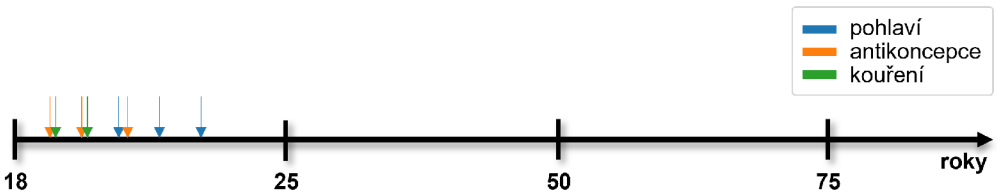

Obr. 7.10: Graf počtů detekovaných driftů v konceptu pro danou detekční metodu a dataset

Nejvíce detekovaných varovných úrovní zaznamenala metoda DDM v datasetu GM na metabolomických modelech predikujících užívání antikoncepce. Na druhou stranu metoda EDDM v případě klasifikace kuřáků nezaznamenala varování ani v jednom datasetu. Závěrem je možno z grafu na obrázku 7.9 konstatovat, že metoda DDM zaznamenala více varovných úrovní než metoda EDDM.

Nejhojnější počet změn zaregistrovala metoda EDDM v predikci metabolomického modelu klasifikujícího pohlaví v datasetu BM. DDM nezaznamenala v případě datasetu BM žádnou změnu v distribuci dat. Na základě grafu na obrázku 7.10 lze vyvodit, že metoda EDDM je vhodnější pro detekci drift konceptu v metabolomických datech, což odpovídá zjištěním ve studii od Schwarzerové a spol. [53]. Zároveň jsou výše popsané výsledky bakalářské práce podpořeny konferenčním příspěvkem [54]. Nicméně by se mohlo zdát, že toto tvrzení vyvrací počet detekovaných změn u modelu klasifikujícího kuřáky v případě detekční metody DDM v datasetu GM. Avšak nutno podotknout, že se potvrdil fakt, jenž je zmíněn v kapitole 4.1 týkající se zde použitých detekčních metod založených na statistické regulaci procesů. A to, že metoda EDDM je nevhodná pro detekci drift konceptu u nevyvážených dat. Tato situace nastala právě u modelů predikujících kuřáky do čtyř tříd, jejichž četnosti byly značně nevyvážené.

V porovnání metod detekce drift konceptu DDM a EDDM na predikčních modelech plyne, že metoda EDDM je daleko citlivější na detekci změn konceptu viz obrázek 7.10. Naopak metoda DDM je citlivější pro detekci varovných úrovní viz obrázek 7.9.

Ve studii od Schwarzerové a spol. [53] je poznamenáno, že jeden zmatoucích faktorů v lidské metabolomické analýze je věk. Proto byla vytvořená časová osa na obrázku 7.11 reprezentující staří pacientů, u kterých byla identifikována změna v konceptu.




Obr. 7.11: Vizualizace detekovaný drift konceptů v závislosti na odhalení matoucího faktoru

Jak můžeme vidět na obrázku 7.11, všichni pacienti, u kterých byla shledána změna v konceptu, jsou ve stáří od 18 do 25 let. Tato doba vypovídá o vývojovém přechodu z dospívání do dospělosti. Tímto práce přináší potvrzení se studií [53] o odhalení matoucího faktoru v metabolomické analýze. Toto zjištění pomůže vytvořit přesnější modely pro včasnou diagnostiku, která je nezbytná pro úplné uzdravení nebo ekonomicky přívětivější léčbu pro pacienty spadající do věkové skupiny 18-25 let.

# Závěr

Tato bakalářská práce se zabývá problematikou strojového učení, a to analýzou drift konceptu v dynamicky se měnících datech. Nejprve byly natrénovány predikční modely založené na algoritmech strojového učení. Predikční modely byly trénovány na koncentracích metabolitů a jejichž cílem bylo zařazení objektů do náležité třídy v rámci daného klasifikačního problému. Klasifikační problémy představovaly: pohlaví, užívání antikoncepce a kouření. Konkrétně byl drift koncept detekován na modelem vytvořené predikci, pokud byla detekčními algoritmy shledána změna v distribuci dat. Tyto algoritmy zaznamenávaly dvě úrovně drift konceptu. První úroveň varovala před případnou změnou v konceptu. Zatímco druhá úroveň signalizovala definitivní změnu v distribuci dat.

Práce nejprve pojednává o samotném drift konceptu. V další kapitole je jedna ze stěžejních částí práce věnovaná metabolomice. Metabolomika přináší novou éru umožňující se věnovat predikci fenotypu na základě koncentrací metabolitů. Třetí kapitola popisuje algoritmy predikčních modelů. Poté následuje literární rešerše zabývající se metodologii detekce a korekce drift konceptu. Některé metody detekce byly následně použity v praktické části, o které pojednává kapitola 7 s názvem Implementace detektorů drift konceptu.

Hlavní cíl bakalářské práce spočíval v detekci drift konceptu pomocí detekčních metod DDM a EDDM v metabolomickém datasetu na různých predikčních modelech. Díky tomu, bakalářská práce přinesla 8 predikčních modelů predikujících daný klasifikační problém na základě koncentrací metabolitů. Dále bylo po zpracování výsledků zjištěno, že metoda EDDM je pro detekci drift konceptu vhodnější než metoda DDM. Na základě nevyváženosti vystupních dat u klasifikačního problému kouření, byl experimentálně ověřen poznatek, že metoda EDDM je nevhodná v případě nevyvážených dat. Na závěr byl odhalen matoucí faktor v metabolomické analýze, kdy bylo zjištěno, že detekované změny v distribuci dat souvisí s věkem pacienta. Hlavní poselství této práce spočívá v odhalení inovativního přístupu v metabolomické analýze s detekcí nežádoucícho faktoru, za který lze považovat koncept drift.

# Literatura


[1] GAMA, Joao; et al. *Learning with Drift Detection*. In: Advances in Artificial Intelligence – SBIA 2004. Ed. by Ana L. C. Bazzan and Sofiane Labidi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 286–295. isbn: 978-3-540-28645-5.

[2] WEBB; Geoffrey I.; et al. *Characterizing concept drift*. Data Mining and Knowledge Discovery, 2016, 30.4: 964-994.

[3] YU, Shujian; et al. *Concept drift detection and adaptation with hierarchical hypothesis testing*. Journal of the Franklin Institute, 2019, 356.5: 3187-3215.

[4] BROWNLEE, Jason. *A gentle introduction to concept drift in machine learning*. Machine Learning Mastery, 2018.

[5] CHEN, Jiaoyan; et al. *Knowledge graph embeddings for dealing with concept drift in machine learning*. Journal of Web Semantics, 2021, 67: 100625.

[6] DE BARROS, Roberto Souto Maior; DE CARVALHO SANTOS, Silas Garrido T. *An overview and comprehensive comparison of ensembles for concept drift*. Information Fusion, 2019, 52: 213-244.

[7] CHANDRA, Swarup; et al. *An adaptive framework for multistream classification*. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016. p. 1181-1190.

[8] IDLE, Jeffrey R.; GONZALEZ, Frank J. *Metabolomics*. Cell metabolism, 2007, 6.5: 348-351.

[9] *An introduction to metabolomics*. (b.r.). EMBL-EBI [online]. Dostupné z: <`https://doi.org/10.6019/tol.mbs.2014.00001.1`>.

[10] ZHANG, Aihua; et al. *Modern analytical techniques in metabolomics analysis*. Analyst, 2012, 137.2: 293-300.

[11] RAJA, P. M. V.; BARRON, A. R. *NMR Spectroscopy*. Rice University.

[12] ARYAL, Sagar. *Nuclear magnetic resonance (NMR) spectroscopy*. 2018.

[13] DE HOFFMANN, E.; STROOBANT, V. *Mass spectrometry: principles and applications*. John Wiley  Sons: 2007.

[14] SODHI, Pinky; AWASTHI, Naman; SHARMA, Vishal. *Introduction to machine learning and its basic application in python.* In: Proceedings of 10th International Conference on Digital Strategies for Organizational Success. 2019.

[15] SMITH, Jennifer. *Researchers tackle bias in algorithms.* JULY 4, 2017.

[16] BROWNLEE, Jason. *Machine learning mastery with Python: understand your data, create accurate models, and work projects end-to-end.* Machine Learning Mastery, 2016.

[17] HOSMER; LEMESHOW, S. *Applied Logistic Regression.* 2nd ed., New York; Chichester, Wiley, 2000, ISBN 0-471-35632-8

[18] ONTIVERO-ORTEGA; Marlis; et al. *Fast Gaussian Naïve Bayes for searchlight classification analysis.* Neuroimage, 2017, 163: 471-479.

[19] BREIMAN, L. *Bagging predictors.* Machine learning, 1996, 24.2: 123-140.

[20] *What is a Random Forest?.* TIBCO [online]. Dostupné z: <`https://www.tibco.com/reference-center/what-is-a-random-forest`>.

[21] ZHANG, Yanru; HAGHANI, Ali. *A gradient boosting method to improve travel time prediction.* Transportation Research Part C: Emerging Technologies, 2015, 58: 308-324.

[22] PESARANGHADER, Ali; VIKTOR, Herna L. *Fast hoeffding drift detection method for evolving data streams.* In: Joint European conference on machine learning and knowledge discovery in databases. Springer, Cham, 2016. p. 96-111.

[23] KNOTEK, Jan; Welma PEREIRA. *Survey on Concept Drift.* MUNI [online]. Dostupné z: <`https://is.muni.cz/el/1433/podzim2011/PA164/um/drift_detection_methods.pdf`>.

[24] ROSS, Gordon J.; et al. *Exponentially weighted moving average charts for detecting concept drift.* Pattern recognition letters, 2012, 33.2: 191-198.

[25] MITTAL, Veena; KASHYAP, Indu. *Online methods of learning in occurrence of concept drift.* International Journal of Computer Applications, 2015, 117.13.

[26] DA COSTA, Fausto G.; et al. *Multidimensional surrogate stability to detect data stream concept drift.* Expert Systems with Applications, 2017, 87: 15-29.

[27] SPINOSA, Eduardo J.; DE LEON F. DE CARVALHO, André Ponce; GAMA, Joao. Olindda: *A cluster-based approach for detecting novelty and concept drift in data streams*. In: Proceedings of the 2007 ACM symposium on Applied computing. 2007. p. 448-452.

[28] DREDZE, Mark; OATES, Tim; PIATKO, Christine. *We're not in kansas anymore: detecting domain changes in streams*. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. 2010. p. 585-595.

[29] SETHI, Tegjyot Singh; KANTARDZIC, Mehmed. *Don't pay for validation: Detecting drifts from unlabeled data using margin density*. Procedia Computer Science, 2015, 53: 103-112.

[30] CHANDRA, Swarup; et al. *An adaptive framework for multistream classification*. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016. p. 1181-1190.

[31] GRULICH, Philipp M.; et al. *Scalable Detection of Concept Drifts on Data Streams with Parallel Adaptive Windowing*. In: EDBT. 2018. p. 477-480.

[32] AWAD, Mariette; KHANNA, Rahul. *Support vector regression*. In: Efficient learning machines. Apress, Berkeley, CA, 2015. p. 67-80.

[33] *Concept-drift-datasets-scikit-multiflow*. Scikit-Multiflow [online]. Dostupné z: <`https://github.com/ogozuacik/concept-drift-datasets-scikit-multiflow`>.

[34] CHU, Xiaojing; et al. *Integration of metabolomics, genomics, and immune phenotypes reveals the causal roles of metabolites in disease*. Genome biology, 2021, 22.1: 1-22.

[35] *Standard Scaler*. Scikit [online]. Dostupné z: <`https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`>.

[36] *PCA*. Scikit [online]. Dostupné z: <`https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`>.

[37] *Select K Best*. Scikit [online]. Dostupné z: <`https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html`>.

[38] *Complete-Feature-Selection*. GitHub [online]. Dostupné z: <`https://github.com/krishnaik06/Complete-Feature-Selection`>.

[39] VAN ROSSUM, G.; DRAKE, F. L. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

[40] MCKINNEY, W.; et al. *Data structures for statistical computing in python*. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

[41] HARRIS, Charles R.; et al. *Array programming with NumPy*. Nature, 2020, 585.7825: 357-362.

[42] HUNTER, J. D. *Matplotlib: A 2D graphics environment*. Computing In Science Engineering 9, 3. 90-95. 2007.

[43] PEDREGOSA, F.; et al. *"Scikit-learn: Machine Learning in Python,"*. Journal of Machine Learning Research, vol. 12, p. 2011.

[44] MONTIEL, Jacob; et al. *Scikit-multiflow: A multi-output streaming framework*. The Journal of Machine Learning Research, 2018, 19.1: 2915-2914.

[45] *Logistic regression*. Scikit [online]. Dostupné z: <`https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`>.

[46] *Gradient Boosting Classifier*. Scikit [online]. Dostupné z: <`https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html`>.

[47] *Random Forest Classifier*. Scikit [online]. Dostupné z: <`https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`>.

[48] *Gaussian NB*. Scikit [online]. Dostupné z: <`https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html`>.

[49] *Cross validation*. Scikit [online]. Dostupné z: <`https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html`>.

[50] ANDRADE, Jonata Jefferson; et al. *PREDICTION OF THE PERFORMANCE OF BITUMINOUS MIXES USING ADAPTIVE NEURO-FUZZY INFERENCE SYSTEMS PREVISÃO DO DESEMPENHO DE MISTURAS BITUMINOSAS USANDO SISTEMAS DE INFERÊNCIA ADAPTATIVA NEURO-DIFUSA*. 2020.

[51] *F1 metric classification report*. Scikit [online]. Dostupné z: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html>.

[52] *Receiver Operating Characteristic curve*. Scikit [online]. Dostupné z: <https://scikit-learn.org/stable/auto_examples/model_selection/plot_det.html#sphx-glr-auto-examples-model-selection-plot-det-py>.

[53] SCHWARZEROVA, Jana; et al. *An Innovative Perspective on Metabolomics Data Analysis in Biomedical Research Using Concept Drift Detection*. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2021. p. 3075-3082.

[54] KOSTOVAL, Ales; et al. *Concept Drift Detection in Prediction Classifiers for Determining Gender in Metabolomics Analysis*. In: Proceedings I of the 28th Conference. STUDENT EEICT 2022. General papers. ISBN 978-80-214-6029-4.

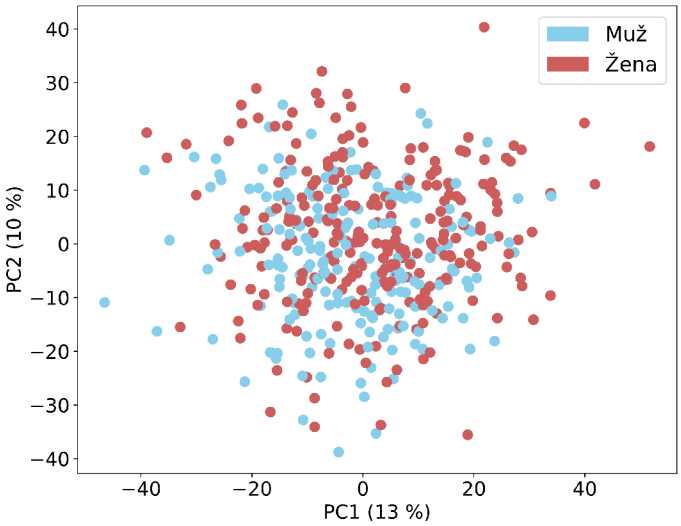
# Seznam symbolů a zkratek

| | |
|---|---|
| **mRNA** | mediátorová ribonukleová kyselina |
| **SNMR** | spektroskopie nukleární magnetické rezonance |
| **TOFMS** | hmotnostní spektrometrie doby letu |
| **LR** | logistická regrese |
| **GB** | gradient boosting |
| **RF** | náhodný les |
| **NB** | Naivní Bayesovský klasifikátor |
| **DDM** | metoda detekce driftu |
| **EDDM** | metoda včasné detekce driftu |
| **PCA** | analýza hlavních komponent |
| **ANOVA** | analýza rozptylu |
| **ROC** | operační charakteristika přijímače |
| **AUC** | plocha pod krivkou |
| $f$ | transformační funkce |
| $X$, $X_i$, $X_n$, $x_t$ | vstupní data |
| $Y$ | výstupní data |
| $\sigma_i$, $\sigma_t$ | směrodatná odchylka |
| $p_i$ | míra chybovosti |
| $p_{min}$ | minimální chyba |
| $\sigma_{min}$ | minimální směrodatná odchylka |
| $\alpha$ | práh |
| $\beta$ | práh |
| $p'_i$ | průměrná vzdálenost mezi chybami |
| $s'_i$ | směrodatná odchylka $p'_i$ |

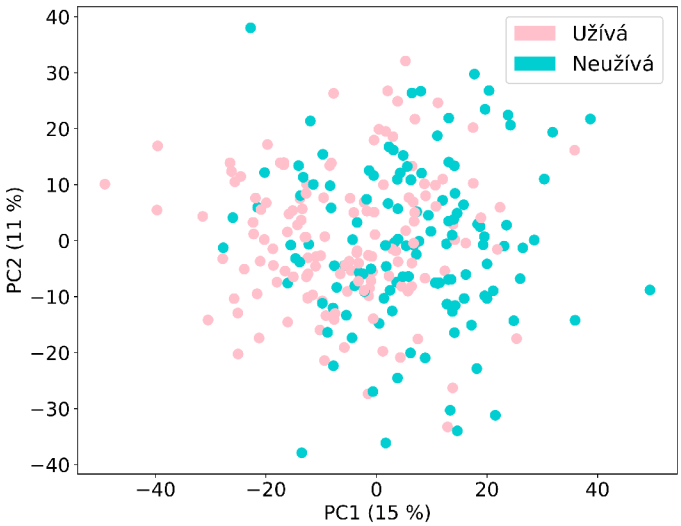
| | |
|---|---|
| $p'_{max}$ | maximální $p'_i$ |
| $s'_{max}$ | maximální $s'_i$ |
| $T(r_0, r_r, n_0, n_r)$ | 4D funkce |
| $r_0$, $r_r$ | počet správných klasifikací v daném případě |
| $W$ | speciální případ klasifikace |
| $n_0$, $n_r$ | počet nesprávných klasifikací v daném případě |
| $p\prime$ | průměrná vzdálenost mezi $r_0$, $r_r$ a $n_0$, $n_r$ |
| $p_t$ | binární parametr |
| $t$ | čas |
| $Z_t$ | název proměnné u vážené regulace procesů |
| $\sigma^2_{Z_t}$ | rozptyl $Z_t$ |
| $p_0$ | pravděpodobnost chybovosti před driftem |
| $p_{0,t}$ | odhad $p_0$ v čase $t$ |
| $X_t$ | vstupní vzorek v čase $t$ |
| $L$ | kontrolní parametr |
| $\lambda$ | koeficient zapomínání |
| $\delta$ | pravděpodobnostní konstanta |
| $\epsilon_H$ | Hoeffdingovo číslo |
| $n$ | počet vstupních prvků |
| $p^1_t$ | pravděpodobnost výskytu 1 za čas $t$ |
| $p^1_{max}$ | maximální pravděpodobnost výskytu 1 |
| $\bar{x}_t$ | průměrný odhad z pozorovaných dat $x_t$ |
| $C_t$ | kumulovaný součet |
| $\lambda_P$ | horní hranice |
| $\lambda_N$ | dolní hranice |

| | |
|---|---|
| $y_i$ | výstupní třída |
| $\xi_i$ | transformační operátor |
| $C$ | konstanta váhy chybné klasifikace |
| $w$ | váha |
| $b$ | práh |
| $f(x, w)$ | aproximační funkce |
| $M$ | řád polynomu |
| $w_i$ | šířka nadroviny |
| $\epsilon$ | ztrátová funkce |
| $B$ | magnetická indukce |
| $f$ | frekvence |

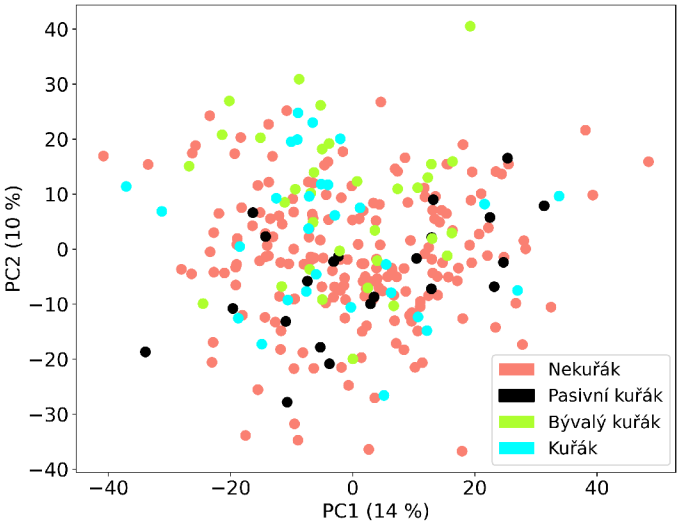
# A Vizualizace PCA GM datasetu




Obr. A.1: Vizualizace analýzy hlavních komponent GM datasetu s vyznačenými očekávanými daty pro klasifikaci pohlaví




Obr. A.2: Vizualizace analýzy hlavních komponent GM datasetu s vyznačenými očekávanými daty pro klasifikaci užívání antikoncepce




Obr. A.3: Vizualizace analýzy hlavních komponent GM datasetu s vyznačenými očekávanými daty pro klasifikaci kuřáků