Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra informačního inženýrství

Diplomová práce

Využití strojového učení pro snížení rizik při užívání psychedelických látek

Bc. Štěpán Pešout

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Provozně ekonomická fakulta

ZADÁNÍ DIPLOMOVÉ PRÁCE

Bc. Štěpán Pešout

Informatika

Název práce

Využití strojového učení pro snížení rizik při užívání psychedelických látek

Název anglicky

Using machine learning to reduce the risks of psychedelic substance use

---

Cíle práce

Cílem práce je vytvoření několika predikčních modelů, jejichž hlavním účelem bude co nejpřesněji předvídat charakter psychedelické zkušenosti konkrétního uživatele na základě jím zadaných údajů. Mezi ně patří zejména demografické informace, druh psychedelické látky, předchozí zkušenosti či předpokládané prostředí, kde se chystá látku užít. Potenciální uživatel tak získá další informace, které mu pomohou učinit rozhodnutí, zda bude chtít zkušenost podstoupit.

Metodika

Teoretická část práce spočívá ve studiu odborných zdrojů, které se týkají zejména algoritmů pro vytváření modelů strojového učení a umělé inteligence, ale také psychedelických látek. Bude stručně shrnuta jejich role v historii a současnosti a také efekty a vlivy na uživatele těchto substancí. Na základě zmíněných podkladů bude možné formulovat teoretická východiska nutná pro další práci.

V rámci praktické části budou zpracována data dlouhodobě získávaná od uživatelů psychedelických látek pomocí speciální mobilní aplikace. Budou převedena do formátu, který je vhodný pro následnou klasifikaci. Na základě nich bude navrženo několik typů predikčních modelů využívajících metody strojového učení a umělé inteligence. Jejich výkonnost bude experimentálně ověřena a kvantifikována. Modely budou navrženy s ohledem na maximalizaci přesnosti, preciznosti a dalších obdobných ukazatelů a budou prezentovány ty, které dosahovaly nejlepších výsledků.

Doporučený rozsah práce

50-60

Klíčová slova

strojové učení, umělá inteligence, predikční model, klasifikace, psychedelické látky, snížení rizik

---

Doporučené zdroje informací

GÉRON, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. Second edition. Beijing: O'Reilly, 2019. ISBN 978-1-492-03264-9.

RUSSELL, Stuart J. a Peter NORVIG. Artificial intelligence: a modern approach. 3rd ed. Harlow: Pearson Education, c2014. ISBN 978-1-29202-420-2.

TAN, Pang-Ning, Michael STEINBACH a Vipin KUMAR. Introduction to data mining. Harlow: Pearson Education, 2013. Pearson new international edition. ISBN 978-1-292-02615-2.

TYLŠ, Filip. Fenomén psychedelie: subjektivní popisy zážitků z experimentální intoxikace psilocybinem doplněné pohledy výzkumníků. Vydání druhé. Praha: Dybbuk, 2020. ISBN 978-80-7438-226-0.

Předběžný termín obhajoby

2022/23 LS – PEF

Vedoucí práce

Ing. Josef Pavlíček, Ph.D.

Garantující pracoviště

Katedra informačního inženýrství

Elektronicky schváleno dne 31. 10. 2022

Ing. Martin Pelikán, Ph.D.

Vedoucí katedry

Elektronicky schváleno dne 28. 11. 2022

doc. Ing. Tomáš Šubrt, Ph.D.

Děkan

V Praze dne 31. 03. 2023

Čestné prohlášení

Prohlašuji, že svou diplomovou práci „Využití strojového učení pro snížení rizik při užívání psychedelických látek“ jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu použitých zdrojů na konci práce. Jako autor uvedené diplomové práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 31. 3. 2023 _____

Poděkování

Rád bych touto cestou poděkoval Ing. Josefu Pavlíčkovi, Ph.D. za vedení práce a vstřícnost při konzultacích; dále také Dominice Dojčánové, MSc. a Michalu Šenkovi, kteří mi pomohli sehnat data z mobilní aplikace iTrip, bez nichž by tato práce nemohla vzniknout.

Využití strojového učení pro snížení rizik při užívání psychedelických látek

Abstrakt

Tato práce hledá možnosti využití strojového učení ke snížení rizik spojených s užíváním psychedelických látek. Problém řeší pomocí několika klasifikačních modelů, které predikují charakter psychedelické zkušenosti.

V teoretické části jsou formulována východiska pro psychedelické zážitky, různé typy látek a jejich využití. Je popsán vliv myšlenkového nastavení a prostředí na průběh zkušenosti a jsou představeny možnosti kvantifikace různých jejích aspektů. Dále jsou zkoumány různé algoritmy strojového učení, tvorba predikčních modelů a optimalizace jejich počáteční konfigurace.

Praktická část práce popisuje nutné úpravy dat, formuluje klasifikační úlohu a identifikuje předpovídané proměnné. Poté je implementován evoluční algoritmus řešící optimalizaci parametrů, popsán způsob výběru atributů a představeno pět různých predikčních modelů. Práce následně vyhodnocuje a srovnává jejich výkonnost.

Získané výsledky jsou analyzovány z různých pohledů. Jsou uvedeny zjištěné výhody a nevýhody implementace evolučního algoritmu. Práce hodnotí naplnění stanovených cílů a představuje možnosti využití modelů v praxi. Taktéž ukazuje, kde je prostor pro další zlepšení a případný rozvoj.

Klíčová slova: strojové učení, umělá inteligence, predikční model, klasifikace, psychedelické látky, snížení rizik

Using machine learning to reduce the risks of psychedelic substance use

Abstract

This work seeks to use machine learning to reduce the risks associated with psychedelic substance use. It addresses the problem using several classification models that predict the nature of the psychedelic experience.

The theoretical part formulates the background of psychedelic experiences, different types of substances and their uses. The influence of mind-set and environment on the course of the experience is described and the possibilities of quantifying different aspects of the experience are presented. Furthermore, various machine learning algorithms, the creation of prediction models and the optimization of their initial configuration are explored.

The practical part of the thesis describes the necessary data modifications, formulates the classification task and identifies the predicted variables. Then, an evolutionary algorithm addressing parameter optimization is implemented, a method for attribute selection is described, and five different prediction models are introduced. The paper then evaluates and compares their performance.

The obtained results are analyzed from different perspectives. The observed advantages and disadvantages of the implementation of the evolutionary algorithm are presented. The thesis evaluates the achievement of the set goals and presents possibilities for the application of the models in practice. It also indicates areas for further improvement and potential development.

Keywords: machine learning, artificial intelligence, prediction model, classification, psychedelics, risk reduction

Obsah

Seznam obrázků

Seznam tabulek

1 Úvod

V současné době zájem lidí o psychedelika vykazuje vzestupný trend. Důvody užívání mohou být různé – od čistě rekreačních, přes spirituální až po léčbu psychických onemocnění. Ačkoli jsou rizika fyzických dopadů užívání a vzniku závislosti je zanedbatelné, je zásadní dbát na správnou souhru myšlenkového nastavení (set) a prostředí, kde se tato zkušenost odehrává (setting). Tyto parametry mají podstatný vliv na průběh zážitku. V případě jejich nevhodné kombinace se může objevit náročná zkušenost, která je typická nepříjemnými či úzkostnými pocity.

Tato práce vychází z obecného předpokladu, že na snižování negativních dopadů činností má velký vliv informovanost. Lidé, kteří znají potenciální nebezpečí, mohou přijmout opatření k jejich předcházení a případně se připravit na eventualitu výskytu těchto komplikací. Rizika užívání psychedelických látek tedy mohou být redukována, pokud potenciální uživatelé budou předem informováni, že tato rizika existují a čeho se týkají.

Charakter psychedelické zkušenosti lze do určité míry zpětně kvantifikovat pomocí dotazníku. Díky tomu je možné měřit a porovnávat jednotlivé aspekty průběhu zážitku – jak negativní, tak i pozitivní. Teoreticky tedy platí, že by mělo být možné na základě dostatečně kvalitních dat ohledně průběhu zážitku a počátečního setu, settingu a dalších informací o uživateli formulovat klasifikační nebo regresní úlohu řešitelnou pomocí metod strojového učení.

Hlavním přínosem této práce bude tedy vytvoření predikčního modelu, který by v případě praktického nasazení pomáhal snižovat rizika a negativní dopady užívání psychedelických látek. V současné době existuje velmi málo akademických publikací, které by popisovaly využití strojového učení pro predikci charakteru této zkušenosti a proto má toto téma poměrně velký potenciál.

2 Cíle práce a metodika

2.1 Cíle práce

Cílem práce je vytvoření několika predikčních modelů, jejichž hlavním účelem bude co nejpřesněji předvídat charakter psychedelické zkušenosti konkrétního uživatele na základě jím zadaných údajů. Mezi ně patří zejména demografické informace, druh psychedelické látky, předchozí zkušenosti či předpokládané prostředí, kde se chystá látku užít. Potenciální uživatel tak získá další informace, které mu pomohou učinit rozhodnutí, zda bude chtít zkušenost podstoupit.

2.2 Metodika

Teoretická část práce spočívá ve studiu odborných zdrojů, které se týkají zejména algoritmů pro vytváření modelů strojového učení a umělé inteligence, ale také psychedelických látek. Bude stručně shrnuta jejich role v historii a současnosti a také efekty a vlivy na uživatele těchto substancí. Na základě zmíněných podkladů bude možné formulovat teoretická východiska nutná pro další práci.

V rámci praktické části budou zpracována data dlouhodobě získávaná od uživatelů psychedelických látek pomocí speciální mobilní aplikace. Budou převedena do formátu, který je vhodný pro následnou klasifikaci. Na základě nich bude navrženo několik typů predikčních modelů využívajících metody strojového učení a umělé inteligence. Jejich výkonnost bude experimentálně ověřena a kvantifikována. Modely budou navrženy s ohledem na maximalizaci přesnosti, preciznosti a dalších obdobných ukazatelů a budou prezentovány ty, které dosahovaly nejlepších výsledků.

3 Teoretická východiska

3.1 Psychedelické látky

Psychedelika je možné definovat jako chemicky různorodé substance, které mění vědomí a jeho kvalitu. Jde o látky rozšiřující vnímání reality tak, že umožňují jedinci rozpoznat skutečnosti, které jsou mimo obvyklý rozsah smyslových zkušeností. Uživatel je schopen překonat běžné myšlenkové vzorce a otevřít nové možnosti sebepoznání. Tyto látky mají potenciál pro změny vědomí, které mohou mít dlouhodobý vliv na jedince nebo jeho hodnotové nastavení. [1][3]

Zejména pro klasická psychedelika (agonisté serotoninového 2A receptoru) je typické, že mají velmi nízký závislostní potenciál. Zároveň u nich téměř chybí riziko předávkování (rozdíl mezi účinnou a letální dávkou je extrémní) a mají nízkou toxicitu. Po jejich požití mohou nicméně nastat nepříznivé psychické reakce. [2][5] Psychedelickým zážitkům a rizikům se dále věnuje kapitola Psychedelická zkušenost.

3.1.1 Klasická psychedelika

Mezi nejužívanější látky patří **LSD** (diethylamid kyseliny lysergové), což je bezbarvá krystalická látka bez chuti. Jedná se o psychedelikum, které ve velmi nízkých (mikrogramových) množstvích vyvolává u uživatele silné halucinogenní účinky. LSD bylo poprvé syntetizováno v roce 1938 ve Švýcarsku doktorem Albertem Hofmannem, který ale jeho psychoaktivní účinky objevil náhodou až v roce 1943. Tato substance od doby svého objevu nalezla mnoho využití v psychiatrii a léčbě duševních chorob. [2][4][5][6]

Dalším klasickým psychedelikem je **psilocybin**. Jedná se o psychoaktivní alkaloid a hlavní psychedelickou složku halucinogenních hub např. z rodu *Psilocybe* (lysohlávka). V současné době je mu věnována velká pozornost vzhledem k jeho potenciálním terapeutickým účinkům a možnostem využití ve výzkumu. Zároveň pravděpodobně jde o nejčastěji užívané (klasické) psychedelikum v České republice, s nímž má zkoušenost 7,3 % populace. Psilocybin byl poprvé izolován a identifikován Albertem Hofmannem v roce 1958. Tuto látku lze připravit i synteticky; je bílá a krystalická. [7][8][9]

Mezi klasická psychedelika patří také **DMT** (dimethyltryptamin), který je velmi často užíván jako součást nápoje **ayahuasca** (původně připravována amazonskými domorodci, ale známá i v západním světě). DMT je produkované rostlinami a přítomné ve stopovém množství i v lidském těle. Dále sem lze zařadit **5-MeO-DMT** (5-methoxy-N,N-dimethyltryptamin) – silné psychedelikum obsažené v sekretu ropuchy coloradské. Klasickými psychedeliky jsou i **meskalin** (tradičně užívaný původními obyvateli Ameriky) nebo **ibogain** – alkaloid přítomný především v tropickém keři iboga. [4][10][11][12][13]

3.1.2 Ostatní psychedelika

Často užívanou látkou je **MDMA** (3,4-methylendioxymethamfetamin), které se označuje jako extáze nebo taneční droga. Jedná se o synteticky připravovanou látku, která patří mezi amfetaminy. MDMA má značné stimulační účinky, takže je uživatel schopen zvýšené fyzické aktivity, která v extrémních případech může vést k úplnému vyčerpání a kolapsu. Po konzumaci se dostavují mírné psychedelické účinky, nárůst empatie a příjemných pocitů obecně. Tato substance získala na popularitě zejména v 80. a 90. letech minulého století. [4]

Primárně jako anestetikum je v současné době v medicíně využíván **ketamin**. V nižších dávkách má psychedelické účinky, pro něž je užíván jak rekreačně, tak i jako účinné antidepresivum. Některé studie také prokázaly jeho pozitivní vliv v léčbě závislostí na alkoholu či jiných drogách. Ketamin krátkodobě zvyšuje krevní tlak, takže jeho užití může být nevhodné u lidí s hypertenzí. [4][14][15]

V prostředí hudebních klubů bývá užíváno **2C-B** (4-brom-2,5-dimethoxyfenylethylamin), což je syntetické psychedelikum připravené Alexanderem Shulginem v roce 1974. V nízkých dávkách působí spíše jako lehký stimulant, zvyšuje náklonnost k lidem a empatii (účinky srovnatelné s MDMA). Halucinogenní působení začíná být patrné až s vyššími dávkami. Vrcholné psychedelické zážitky nejsou typickým jevem. [4][16]

Některé zdroje řadí mezi psychedelika také **salvinorin A**, tedy hlavní psychoaktivní látku šalvěje divotvorné. Chemická struktura je velmi odlišná od všech klasických psychedelik a bývá považován za nejsilnější přírodní psychoaktivní látku. [4][17][18]

Pro své psychotropní a halucinogenní účinky je někdy užíván také **muscimol**, tedy hlavní alkaloid obsažený v muchomůrce červené (a příbuzných druzích). [19]

Určité psychedelické účinky mohou prožívat i uživatelé **konopí** (pro jeho obsah kanabinoidů). Konopí je ale významné zejména v kombinaci s jinými látkami (zejména klasickými psychedeliky), protože může značně modifikovat průběh zkušenosti a to jak pozitivním, tak i negativním způsobem. [4][20][21]

3.1.3 Využití psychedelik v lékařství

Významné využití nalézají psychedelika v léčbě pacientů s depresí. Mezi látky, které se v poslední době zkoumají nejčastěji patří psilocybin a ketamin. V posledních letech byla provedena řada klinických studií, které naznačují, že psychedelická či psychedeliky asistovaná terapie může být účinná u lidí trpících i středně těžkou až těžkou depresí, u kterých selhala běžná léčba. [7][22][23]

Psilocybin se podle několika klinických studií ukázal být účinným při léčbě deprese. Pacienti se v průměru cítí méně deprimovaní a úzkostní po užití této látky, přičemž tato účinnost může trvat i několik měsíců po jediné dávce. Také se ukázalo, že psilocybin v kombinaci s psychoterapií vedl k významnému snížení symptomů deprese a úzkosti u pacientů s pokročilým stádiem rakoviny, což ho potenciálně předurčuje také k využití v paliativní péči. [7][22][23]

Ketamin je v současnosti v medicíně využívaný jako anestetikum, ale může být užitečný i v léčbě deprese, zejména u pacientů, kteří neodpovídají na standardní terapii. V tomto ohledu se ukázal jako rychlý a účinný terapeutický prostředek s vysokou mírou odpovědi a remise v krátkodobém horizontu, ale z dlouhodobého hlediska jsou výsledky zatím rozporuplné. Mechanismus účinku ketaminu v léčbě deprese zatím není přesně známý. Na rozdíl od psilocybinu je ale již používán v klinické praxi díky tomu, že je v mnoha zemích schváleno jeho použití jako léčivo. [24][25]

Psychedelické látky se stávají stále více zkoumanými jako prostředek pro léčbu závislostí. Studie naznačují, že mohou poskytnout pacientům novou perspektivu, zvýšenou motivaci k změně a pomáhají zmírnit abstinenční příznaky. Ukazuje se, že psychedelika mají potenciál v léčbě závislosti na alkoholu, tabáku, opiátech nebo kokainu. Za tímto

účelem jsou zkoumána zejména klasická psychedelika a ketamin. Studie často uvádí, že ke zlepšením došlo již po jedné dávce zkoumané látky. [15][23][26][27]

3.1.4 Nelékařská využití psychedelik

Náboženský či rituální kontext užití je v některých církvích či šamanských kulturách tradiční záležitostí. V současné době ale roste nová vlna náboženského využití psychedelik, která využívá mnoha různých látek (včetně syntetických). Tyto látky mohou pomoci jedincům dosáhnout transcendentních nebo spirituálních prožitků a přinést jim pocit jednoty s duchovním světem. [28][29]

Psychedelika bývají také užívána za účelem osobního rozvoje. Tyto látky mohou pomoci jednotlivci přehodnotit své životní cíle, posílit smysl pro spiritualitu a podpořit kreativitu. Uživatelé často očekávají zlepšení vztahů a emočního zdraví nebo novou perspektivu na svůj život a svět kolem sebe. Mezi populární formy seberozvojového využití psychedelik patří také tzv. microdosing, což je užívání velmi malých dávek psychedelických látek s cílem zlepšit produktivitu a kreativitu. Reálné přínosy microdosingu pro uživatele jsou ale nejasné, protože není zatím podpořen dostatečným množstvím vědeckých důkazů. [30][31]

Mnoho lidí také užívá psychedelika čistě za účelem zábavy a získání zážitku. Často uváděné důvody jsou také touha po relaxaci a zlepšení nálady. [4][32]

3.2 Psychedelická zkušenost

Psychedelická zkušenost je změněný stav vědomí, který je doprovázen vizuálním zkreslením reality a hlubokými smyslovými prožitky. Obvyklý je také pocit ztráty vlastní identity. Jedinec tak často rozpoznává skutečnosti, které se nachází mimo obvyklý rozsah smyslových zkušeností. Obvyklé jsou pseudohalucinace, tedy představy a fantazie, které se liší od pravých halucinací zejména v tom, že si je osoba vědoma, že jde o jevy, které nejsou součástí vnějšího světa. Uživatelé v některých případech dokonce řadí psychedelické zkušenosti mezi nejvýznamnější zážitky svého života. [9][34]

Pro takovýto zážitek je typické, že jeho charakter je možné predikovat poměrně obtížně; bývá pokaždé odlišný a jedinečný. Kromě druhu, dávky a kvality psychoaktivní

látky totiž značně záleží také na dalších faktorech. Podstatné jsou předchozí zkušenosti daného jedince, jeho psychický stav a individuální představy či očekávání, které má před zahájením zážitku. Velmi důležité je také místo a prostředí, kde je látka užita. Často používaným termínem je „set a setting“ označující kombinaci aktuálního stavu psychiky (set) a prostředí v době užití látky (setting). [33][34]

Psychedelika (zejména klasická) se vyznačují velmi nízkou toxicitou, což rapidně snižuje riziko předávkování. Fyzické dopady užívání těchto látek jsou tedy zanedbatelné (kromě případů nevědomé záměny s jinou látkou). Samotná psychedelická zkušenost může být ale v některých případech subjektivně vnímána negativně. Psychedelika pomáhají odhalit skryté emoce či vzpomínky a vyvolávají silné psychické reakce. V kombinaci s nevhodně zvoleným settingem a aktuálním setem může někdy jít o poměrně nepříjemné zážitky, které se mohou projevovat například pocitem zmatenosti nebo paranoie. Pozitivní a negativní prožitky se v průběhu zkušenosti mnohdy střídají. [33][35][36]

Taková psychedelická zkušenost, která je po delší čas doprovázena nepříjemnými pocity se označuje jako náročná. V extrémních případech lze použít i anglický výraz „bad trip“. Mnohdy uživatel ale zpětně vyhodnotí i náročnou zkušenost jako transformativní a přínosnou pro jeho seberozvoj. K integraci takovéhoto zážitku může významně přispět psychoterapie. [35][36]

3.2.1 Kvantifikace charakteru zkušenosti

Existuje vícero způsobů, jak lze za použití dotazníku kvantifikovat subjektivní prožitky lidí, kteří prošli psychedelickou zkušeností. Často využívaná je pětidimenzionální stupnice hodnocení změněných stavů vědomí (5-Dimensional Altered States of Consciousness Rating Scale, zkráceně 5D-ASC). Jde konkrétně o tyto dimenze:

- oceánská bezbřehost (oceanic boundlessness – OBN),
- úzkost spojená s mizením ega (dread of ego dissolution – DED),
- vizuální restrukturalizace (visionary restructuralization – VRS),
- zvukové změny (auditory alterations – AUA),
- pokles bdělosti (vigilance reduction – VIR). [37][38]

Numerické hodnoty pro každou z dimenzí lze získat pomocí reakcí na každý ze série výroků, které popisují různé stavy vědomí. Jedinec označí na škále, jak moc změněný byl stav vědomí v průběhu zkušenosti podle každého z těchto hledisek oproti normálnímu bdělému stavu. [37][38]

Původní návrh této stupnice obsahoval pouze 3 dimenze (OBN, DED, VRS) a pro popis psychedelické zkušenosti se mnohdy používá v této podobě. Nedávné studie také k tomuto účelu využívají rozšířený model obsahující 11 subškál, které představují další dělení těchto 3 dimenzí. [37][38][39]

Důležitým hlediskem, ze kterého lze popsat proběhlou zkušenost je také míra rozpuštění ega. Jde o stav změněného vědomí, při kterém se jedinec cítí, jako by se jeho vnímání „roztékalo“ nebo „rozplynulo“ do okolního prostoru a přestává mít pocit oddělenosti od světa. Ke kvantifikaci lze využít škálu nazvanou Ego-Dissolution Inventory (EDI). Podobně jako u 5D-ASC je možné hodnotu EDI získat pomocí reakcí na tvrzení, která popisují různé stavy vědomí. [40]

Existuje celá řada dalších nástrojů, které slouží k měření různých aspektů změněného stavu vědomí, které hodnotí zkušenost například z pohledu mystiky nebo se zaměřují na existenciální otázky. [41]

3.2.2 Oceánská bezbřehost (OBN)

Oceánská bezbřehost označuje pocit jednoty či splynutí s univerzem spojený s různými spirituálními zážitky. Zahrnuje pozitivní emoce spojené s rozpuštěním vlastního já (ega). Pokud je v této dimenzi skóre vysoké, zkušenost pravděpodobně měla velký terapeutický přínos. Hodnotu OBN lze získat pomocí ohodnocení následujících tvrzení:

1. zažil/a jsem všeobjímající lásku;
2. moje tělesné pocity byly velice příjemné;
3. zažil/a jsem pocit bezbřehé radosti;
4. zdálo se, že konflikty a rozpory se rozplynuly;
5. zdálo se mi, že se vše sjednocuje v jedno;
6. můj prožitek měl i náboženský rozměr;

7. zažil/a jsem něco jako silný úžas;
8. cítil/a jsem, že mám mimořádné schopnosti;
9. připadalo mi, jako bych se vznášel;
10. cítil/a jsem se sjednocen/a se svým okolím;
11. měl/a jsem pocit, že jsem mimo své tělo;
12. cítil/a jsem vše velmi intenzivně;
13. měl/a jsem pocit, jako bych už neměl/a tělo;
14. cítil/a jsem se naprosto volný/á a zproštěný/á veškerých povinností;
15. cítil/a jsem se být propojen/á s vyšší mocí;
16. najednou jsem pochopil/a souvislosti, které mě předtím mátly;
17. starosti a úzkosti každodenního života se zdály nepodstatné;
18. zdálo se, že hranice mezi mnou a mým okolím se stírá;
19. svět se zdál mimo dobro a zlo;
20. měl/a jsem velice originální myšlenky;
21. prožíval jsem minulost, současnost a budoucnost jako jedno;
22. prožil/a jsem hluboký vnitřní klid;
23. moje vnímání času a prostoru bylo změněné, jako kdybych snil/a;
24. všechno kolem mě se zdálo být živoucí;
25. spousta věcí mi připadala neskonale krásná;
26. zažil/a jsem dotek věčnosti;
27. cítil/a jsem, že jsem byl/a v nádherném jiném světě. [37][38][42]

3.2.3 Úzkost spojená s mizením ega (DED)

Mizení nebo rozpuštění ega je termín, který popisuje stav, kdy jedinec cítí, jako by se jeho já (ego) stávalo součástí okolního prostoru. Někdy se dostavuje pocit ztráty či opuštění vlastního těla. V této situaci se jedinec cítí, jako by přestal být odděleným jedincem a stal se součástí něčeho mnohem většího. Jedná se o poměrně silný zážitek a pro jeho pozitivně vnímaný průběh je nutné se oprostit od vazby na sebe sama. Selhání v tomto směru může mít za následek negativní zážitky spojené s pocity ztráty kontroly a s úzkostmi. Náročné zkušenosti jsou tedy zpravidla ty, které mají vysoké DED skóre. Tuto hodnotu lze získat pomocí ohodnocení následujících tvrzení:

1. bál/a jsem se, že nad sebou ztrácím kontrolu;
2. bál/a jsem se, že stav, ve kterém jsem se nacházel/a, bude trvat navždy;
3. moje tělo se cítilo strnulé, bez života nebo jako cizí;
4. cítil/a jsem, jako by mne ovládly temné síly;
5. zažil/a jsem nesnesitelnou prázdnotu;
6. všechno jsem prožíval/a děsivě zkreslené;
7. všechno utíkalo tak rychle, že jsem to nestačil/a sledovat;
8. cítil/a jsem se izolovaný/á od všeho a od všech;
9. cítil/a jsem se jako loutka nebo panenka;
10. cítil/a jsem se, jako bych byl paralyzovaný/á;
11. měl/a jsem pocit, že se stane něco hrozného;
12. cítil/a jsem se ohrožen/a;
13. měl/a jsem pocit, že už vůbec nemám vlastní vůli;
14. po delší časový úsek jsem zůstal/a ustrnulý/á ve velice nepřirozené pozici;
15. cítil/a jsem se neschopný/á udělat sebemenší rozhodnutí;
16. byl/a jsem vyděšený/á, aniž bych přesně věděl/a proč;
17. své okolí jsem zažíval jako divné a zvláštní;
18. čas ubíhal trýznivě pomalu;
19. cítil/a jsem se ztrápený/á;
20. nebyl/a jsem schopen/na dokončit myšlenku, moje myšlení bylo opakovaně nesouvislé;
21. měl/a jsem problém odlišit podstatné věci od nepodstatných. [37][38][42][43]

3.2.4 Vizuální restrukturalizace (VRS)

Vizuální restrukturalizace označuje změněné vnímání zejména z hlediska vizuálních vjemů, různých druhů halucinací apod. Může jít o světelné či barevné záblesky, ale i o složité a komplexní vize a představy. Tato dimenze také reflektuje dojmy, že zvuky ovlivňují barvy či tvary pozorovaných objektů, tedy například pocity vnímání sluchových podnětů pomocí zraku. Na hodnotu VRS mají vliv ohodnocení následujících tvrzení:

1. se zavřenýma očima nebo v absolutní tmě jsem viděl/a barvy;
2. některé každodenní záležitosti získaly zvláštní význam;
3. moje představivost byla extrémně živá;
4. v absolutní tmě nebo se zavřenýma očima jsem viděl/a světla nebo záblesky;
5. na mysl mi přišly věci, o kterých jsem si myslel/a, že jsou již dlouho zapomenuté;
6. okolní předměty mě emočně upoutávaly více než obvykle;
7. byl/a jsem schopný/á si vybavit určité události s extrémní jasností;
8. v absolutní tmě nebo se zavřenýma očima jsem viděl/a pravidelné obrazce;
9. viděl/a jsem věci, o kterých jsem věděl/a že nejsou skutečné;
10. se zavřenýma očima nebo v absolutní tmě se mi promítaly různé scény;
11. mohl/a jsem neobyčejně jasně vidět obrazy ze svých vzpomínek a představ;
12. zdálo se, že barvy věcí se mění podle zvuků a hluku;
13. zdálo se, že tvary se mění podle zvuků;
14. zdálo se, že zvuky ovlivňují to co vidím;
15. mnoho věcí mi připadalo neuvěřitelně legračních;
16. věci v mém okolí se mi zdály menší nebo větší;
17. věci kolem pro mě dostaly nový zvláštní význam. [37][38][42]

3.2.5 Míra rozpuštění ega (EDI)

Koncept rozpuštění ega, tedy pocit rozplynutí hranic těla a osobnosti, byl již popsán v části Úzkost spojená s mizením ega. Pokud je dosaženo vysoké skóre, šlo pravděpodobně o významnou či transformativní zkušenost. V takových případech mnohdy dochází k prohloubení introspekce, k většímu porozumění sobě samému nebo ke změnám ve vztahu k okolnímu světu. Tyto zkušenosti mohou mít dlouhodobý dopad na život jedince. Na rozdíl od hledisek OBN a DED, které jsou součástí pětidimenzionální stupnice hodnocení změněných stavů vědomí, nelze z hodnoty EDI určit, zda se jednalo o příjemnou či náročnou zkušenost. Míru rozpuštění ega lze získat pomocí ohodnocení následujících tvrzení:

1. zažil/a jsem rozpuštění sebe sama nebo svého ega;
2. cítil/a jsem se sjednocen s vesmírem;
3. měl/a jsem pocit sjednocení s ostatními;
4. zažil/a jsem snížení pocitu vlastní důležitosti;
5. zažil/a jsem rozpad sebe sama nebo svého ega;
6. cítil/a jsem se mnohem méně zaujatý/á svými vlastními záležitostmi a starostmi;
7. zcela jsem ztratil/a ponětí o svém egu;
8. veškeré ponětí o sobě a vlastní identitě se rozplynulo. [43][40]

Původní článek, který představil dotazník pro zjišťování míry rozpuštění ega, doporučuje zároveň zavést také dalších 8 tvrzení, která popisují zbytnění ega (ego inflation). Tento jev se však (zejména u klasických) psychedelik v podstatě nevyskytuje a v dotazníku má pouze kontrolní funkci. Práce s hodnotami míry zbytnění ega nedává tedy příliš velký smysl. [40]

3.3 Řešení problémů pomocí strojového učení

Strojové učení je jednou z podoblastí umělé inteligence. Je to obor, který se zabývá návrhem a implementací algoritmů, které umožňují počítačovému systému, aby se učil z předchozích dat. Cílem je vytvoření modelů, které budou schopné na základě dříve vložených trénovacích dat co nejpřesněji předvídat budoucí hodnoty pro nová, zatím neznámá data. Tyto algoritmy se tedy snaží najít v trénovacím souboru různé vzory a vztahy, na základě kterých lze vytvářet predikce. Pro modely strojového učení je typické, že s narůstajícím množstvím dat (předchozích zkušeností) zvyšují svou výkonnost (přesnost předpovědí). [44][45]

Algoritmy je možné rozdělit na několik základních typů, podle toho, jakým způsobem se učí. Jedním z nich je **učení s učitelem** (supervised learning), což je přístup, kdy jsou pro všechna trénovací data poskytnuty také správné výstupy. Dalším typem je **učení bez učitele** (unsupervised learning), tedy technika, kdy výstupy nejsou známé a cílem algoritmu je zejména adaptace na strukturu vstupních dat. Tyto přístupy lze však v některých případech kombinovat (semi-supervised learning), což v praxi znamená, že správné výstupy jsou k dipozici pouze u části tréninkového souboru. [44][45]

Některé algoritmy se vyznačují tím, že je před začátkem procesu učení nutné znát všechna data. Poté je predikční model vytvořen a už není možné, aby zohledňoval nové vstupy a měnil svou strukturu. Takový druh algoritmu je označován jako **dávkový**. V opačném případě, kdy je dodatečná adaptace možná, jde o **inkrementální** algoritmy. [44][45]

Aby bylo možné řešit problém pomocí strojového učení, je nutné jej správně formulovat. Typickými úlohami, které jsou řešitelné pomocí těchto algoritmů jsou **klasifikace** (rozdělení dat do několika předem známých tříd), **regrese** (odhad číselné hodnoty na základě vstupu) nebo **shluková analýza** (zařazení vstupních objektů do skupin, které nejsou předem známé). Kromě těchto nejběžnějších úloh existují i další, které je možné řešit pomocí strojového učení. Výstupem může být například určitá struktura jako je matice nebo graf. [44][45]

3.3.1 Klasifikace

Jak již bylo zmíněno, klasifikační problémy jsou jedním z typů úloh, které je vhodné řešit pomocí strojového učení. Klasifikační algoritmy se zabývají přiřazováním kategorií (tříd) novým datům na základě předchozích vzorů, přičemž těchto tříd je obvykle nízký počet a jsou předem známy. [44][45][46]

Příkladem klasifikační úlohy může být rozdělování žadatelů o půjčku do několika skupin podle jejich solventnosti na základě jejich finanční historie. Dalším možným využitím je detekce spamu v e-mailu (na základě obsahu, odesílatele či jiných dat je rozpoznáváno, zda se jedná o spam či nikoli – přiřazuje se tedy do dvou tříd). V případě, kdy jsou tyto kategorie pouze dvě, se úloha označuje jako binárně klasifikační. [44][45][46]

3.3.2 Vyhodnocování výkonnosti u klasifikačních úloh

Ověření výkonnosti predikčních modelů je důležitým krokem v procesu strojového učení, neboť poskytuje nezbytné informace o tom, jak přesné jsou předpovědi v praxi. Existuje vícero přístupů, jak lze výkonnost měřit, ale všechny jsou založené na tom, že je nutné striktně rozdělit vstupní soubor na trénovací a testovací množinu. Zatímco data v trénovací množině algoritmus použije k nalezení vzorů a vztahů klíčových pro

zařazení do správné třídy, testovací část slouží k následnému zjištění jeho schopnosti predikovat na nových, zatím neznámých datech. [45][46]

V některých případech je vstupní soubor rozdělen na tři části. Kromě trénovací a testovací množiny je oddělena ještě validační část, pomocí které je možné získávat informace o výkonnosti modelu již v průběhu jeho učení na trénovacích datech. Tento přístup je využíván zejména v případech, kdy je proces učení výpočetně náročný, tedy například u umělých neuronových sítí. [45][46][47]

Kromě testovací množiny je vhodné přesnost predikcí vyhodnotit i pro vstupy náležící do trénovacího souboru. Takto je možné eliminovat problém přílišného přizpůsobení modelu trénovacím datům označovaný jako **overfitting**. Tento stav je možné rozpoznat podle toho, že výkonnost modelu je značně vyšší pro trénovací množinu, než pro testovací část dat. [45][46]

Pro přehledné znázornění a vyhodnocení výkonnosti klasifikačního modelu pro testovaný soubor se používá **matice záměn**. V tabulce 1 je uveden příklad pro klasifikaci do tří kategorií, ale tuto matici lze sestavit pro libovolný počet tříd. Tento nástroj umožňuje porovnat předpovězené třídy se skutečností. Na hlavní diagonále jsou umístěny počty správně klasifikovaných vstupů pro každou z kategorií. Ostatní hodnoty představují chybné předpovědi. [45][46]

		Predikovaná třída		
Skutečná třída		třída $= A$	třída $= B$	třída $= C$
	třída $= A$	a_a	a_b	a_c
	třída $= B$	b_a	b_b	b_c
	třída $= C$	c_a	c_b	c_c

Tabulka 1: Matice záměn pro klasifikaci do tří tříd

Zdroj: [46], vlastní zpracování

Matice záměn může být využita pro výpočet různých metrik, které kvantifikují výkonnost modelu. Jednou z nejpoužívanějších je **přesnost** (accuracy – rovnice 1), která vyjadřuje poměr správně klasifikovaných prvků vůči všem prvkům (podíl součtu hodnot na hlavní diagonále k součtu všech hodnot v matici). [46]

$$accuracy = \frac{a_a + b_b + c_c}{a_a + a_b + a_c + b_a + b_b + b_c + c_a + c_b + c_c} \tag{1}$$

Přesnost ale může přinést několik potenciálních problémů. Pokud je dataset nevyvážený (jedna je třída zastoupena více než druhá), může být přesnost klamavá – pokud je například v binární klasifikaci třída A zastoupena v 98 % případů a třída B ve 2 %, pak by klasifikátor předpovídající ve všech případech příslušnost ke kategorii A dosahoval přesnosti 98 %, přestože by v praxi šlo o nepoužitelný model. Dalším limitem je neschopnost rozlišit mezi chybami prvního a druhého typu. Při používání metriky přesnosti je tedy nutné brát ohled na tato omezení. [46]

Pokud klasifikační problém pracuje s více než dvěma třídami a zároveň platí, že tyto kategorie jsou ordinální (lze určit jejich pořadí) je možné použít **rozšířenou přesnost** (extended accuracy – rovnice 2). Tato metrika považuje za korektní jak správné předpovědi, tak i ty, které jsou vzdáleny od správné třídy o jedna (podíl součtu hodnot na hlavní diagonále a dvou sousedních diagonálách k součtu všech hodnot v matici). Nejde o široce využívaný koncept, ale zjištěnou hodnotu je vhodné uvést jako doplňkovou informaci u modelů, které mají nízkou přesnost z důvodu obecně špatné predikovatelnosti příslušnosti ke třídě vycházející z charakteru řešeného problému. [48][49]

$$extended\ accuracy = \frac{a_a + b_b + c_c + a_b + b_c + b_a + c_b}{a_a + a_b + a_c + b_a + b_b + b_c + c_a + c_b + c_c} \tag{2}$$

Další hojně využívanou metrikou je **preciznost** (precision). Využívá se zejména u binárních klasifikačních modelů, kde je definována jako poměr pravdivě pozitivních případů k celkovému počtu případů klasifikovaných jako pozitivní. Preciznost lze zobecnit i pro více tříd. V takovém případě je nutné tuto hodnotu vypočítat pro každou třídu zvlášť a z výsledků následně spočítat aritmetický průměr. Pokud jsou třídy nerovnoměrně zastoupeny, je vhodnější zvolit vážený průměr, který zohlední relativní počty prvků v každé třídě. Rovnice 3 definuje výpočet preciznosti pro tři vyvážené třídy v kontextu tabulky 1. [46][50]

$$precision = \left(\frac{a_a}{a_a + b_a + c_a} + \frac{b_b}{a_b + b_b + c_b} + \frac{c_c}{a_c + b_c + c_c} \right) \ / \ 3 \tag{3}$$

Mezi základní metriky patří také **úplnost** (recall). Používá se nejčastěji s binární klasifikací, kde je možné jí definovat jako podíl správně klasifikovaných pozitivních případů k celkovému počtu skutečně pozitivních případů. Úplnost je také možné využít i pro klasifikaci do více tříd, přičemž je potřeba vypočítat tuto hodnotu pro každou třídu zvlášť. Pro získání finálního údaje je taktéž možné využít aritmetický průměr (nebo vážený, pokud jsou třídy nerovnoměrně zastoupeny). Hlavní rozdíl úplnosti oproti preciznosti je ten, že preciznost se zaměřuje na minimalizaci počtu falešně pozitivních případů a úplnost akcentuje minimalizaci počtu falešně negativních případů. Výpočet úplnosti pro tři vyvážené třídy (v kontextu tabulky 1) je definován rovnicí 4. [46][50]

$$recall = \left(\frac{a_a}{a_a + a_b + a_c} + \frac{b_b}{b_a + b_b + b_c} + \frac{c_c}{c_a + c_b + c_c} \right) \ / \ 3 \tag{4}$$

Kombinací preciznosti a úplnosti je **F-míra** (F-measure), která je definována jako harmonický průměr těchto dvou metrik (rovnice 5). Slouží pro celkové hodnocení výkonu klasifikačního modelu. Její použití je vhodné zejména v případech kdy je optimalizována preciznost a úplnost zároveň a cílem je nalézt vyvážený poměr mezi nimi. F-míra je citlivá na rovnoměrné zastoupení tříd v souboru a její použití pro nevyvážené datasety proto nemusí být vhodné. [46]

$$F\text{-}measure = \frac{2 * precision * recall}{precision + recall} \tag{5}$$

Kromě prostého rozdělení vstupního souboru na trénovací a testovací část je možné pro evaluaci výkonnosti použít **křížovou validaci** (cross-validation). Tato metoda pracuje s předem definovanou hodnotou K. Vstupní data jsou rozdělena na K podmnožin. Jedna podmnožina slouží jako testovací, přičemž je zbytek dat použit pro natrénování

modelu. Tento proces se opakuje přesně K-krát (pro každou z podmnožin) a ve všech případech jsou vypočítány vybrané metriky. Výkonnost modelu je pak možné zjistit pomocí aritmetického průměru z výsledků vypočítaných hodnot pro jednotlivá opakování učícího procesu. Výhodou tohoto přístupu jsou přesnější hodnoty jednotlivých metrik. Nedostatkem je nutnost vytvářet a trénovat K různých modelů, což může být velmi výpočetně náročné. [45][46][47]

3.3.3 Předzpracování dat z hlediska řádků (záznamů)

Aby bylo dosaženo co nejlepších výsledků, je důležité vstupní datový soubor předem analyzovat a předzpracovat. Pokud by byl pro učení použit původní dataset bez jakýchkoli úprav, je velmi pravděpodobné, že to bude mít zásadní negativní vliv na výkonnost modelu. Některé algoritmy strojového učení neumožňují zpracování určitých typů proměnných, takže v extrémním případě by vůbec nebylo možné model vytvořit. [45][46]

V datových souborech mnohdy chybí některé hodnoty. Existuje vícero způsobů, jak tento problém řešit. Nejjednodušším způsobem je smazání řádků (záznamů), ve kterých některé údaje chybí. Nevýhodou tohoto přístupu je potenciální ztráta mnoha informací. Ve většině případů je nejlepším řešením odstranění pouze řádků s relativně velkým množstvím chybějících hodnot. Pokud v záznamu nejsou uvedeny hodnoty pro numerické spojité atributy, je možné je nahradit pomocí mediánu či průměru pro daný sloupec. Chybějící údaje pro kategorické atributy lze nahradit většinově zastoupenou kategorií. Pokročilejším způsobem je sestavení predikčního modelu, který předpovídá hodnotu chybějící proměnné na základě ostatních údajů. [51]

Před započetím učení je důležité ověřit, zda jsou data rovnoměrně rozložena mezi jednotlivé třídy. V případě, že by byl vstupní soubor nevyvážený, algoritmus může přiřazovat vyšší váhu dominantní třídě a špatně tak klasifikovat příslušnost k menšinovým kategoriím. Pokud není možné získat další tréninková data patřící menšinovým třídám, existuje několik přístupů, jak tento problém řešit. V datasetech, které mají velké množství záznamů (více, než je potřeba pro sestavení kvalitního klasifikačního modelu), lze odstranit některé vzorky příslušící k většinové třídě. Opačným přístupem je převzorkování záznamů s menšinovými kategoriemi (prostá duplikace nebo algorit-

mické generování nových dat na základě struktury již existujících záznamů). Některé typy algoritmů strojového učení umožňují deklarování vah pro každou z kategorií, díky čemuž je možné zohlednit nevyváženost datového souboru. [52][53]

Pokud je to možné, je vhodné identifikovat hodnoty, které jsou v kontextu klasifikačního problému logicky nesmyslné. Zejména u dat, která pocházejí z dotazníkových šetření se stává, že některé uvedené záznamy reprezentují stav, který by v reálném světě byl krajně nepravděpodobný či přímo nemožný. Je vhodné zvážit odstranění takovýchto řádků. V praxi však může být jejich identifikace velmi složitá. [46]

3.3.4 Předzpracování dat z hlediska sloupců (atributů)

Některé algoritmy strojového učení vyžadují, aby byly všechny atributy numerické. V ostatních případech číselný typ atributů často pomáhá zvýšit přesnost predikcí. Pokud v kategorické proměnné lze nalézt logické pořadí mezi prvky množiny všech možných hodnot proměnné, převod na numerické hodnoty lze provést prostým nahrazením za čísla odpovídající tomuto pořadí. V případě, že kategorická proměnná nemá ordinální charakter, je možné problém řešit jejím rozdělením. V tabulce 2 je příklad datasetu s jednou kategorickou neordinální proměnou, která byla nahrazena třemi novými tak, aby byly všechny proměnné v souboru numerické – tabulka 3. Jednoduché nahrazení jednotlivých kategorií za čísla v rámci jedné proměnné zde nelze provést, protože by mezi hodnotami bylo uměle vytvořeno pořadí, které v původních datech nebylo přítomno. [46][54]

#	atribut
1	c
2	a
3	b
4	a

Tabulka 2: Soubor s jednou kategorickou proměnnou

Zdroj: [46][54], vlastní zpracování

#	atribut_a	atribut_b	atribut_c
1	0	0	1
2	1	0	0
3	0	1	0
4	1	0	0

Tabulka 3: Soubor se třemi novými numerickými proměnnými

Zdroj: [46][54], vlastní zpracování

U jednotlivých proměnných je vhodné zajistit, aby měly podobné rozsahy hodnot. Velké rozdíly v absolutních hodnotách mohou za použití některých algoritmů (zejména neuronových sítí) způsobovat, že mají proměnné s velkým rozsahem hodnot větší váhu, přestože nejsou důležitější než ostatní. Tato normalizace se často provádí tak, aby všechny hodnoty proměnných ležely mezi 0 a 1. [47][55]

Důležitou úpravou je snižování počtu atributů. Je nutné zachovat ty, které jsou pro klasifikaci nejrelevantnější. Tato redukce má pozitivní vliv na trénování (snižuje výpočetní náročnost) a může zvýšit výslednou výkonnost modelu či následnou interpretovatelnost výsledků. Kromě využití pokročilejších technik selekce příznaků je vhodné také zvážit odstranění atributů, jejichž hodnota v jednotlivých záznamech v souboru často chybí. [46][56][57]

Mezi nejčastější techniky výběru atributů patří filtrace, která využívá statistickou analýzu dat (např. Pearsonův korelační koeficient nebo vzájemná informace). Výhodou filtrace je, že nevyžaduje trénování celého modelu, ale pouze hodnotí významnost jednotlivých atributů. [46][56][57]

Relevantní atributy je také možné vybírat pomocí techniky obalování. Cílem je najít nejlepší kombinaci atributů pomocí trénování nového modelu pro každou z těchto kombinací. Na základě výkonnosti jednotlivých vytvořených modelů je vybrána nejlepší podmnožina atributů. Tento přístup je poměrně spolehlivý a má dobré výsledky, ale při větším počtu atributů může být výpočetní složitost extrémní. [46][56][57]

Techniky filtrace a obalování je možné kombinovat. Tento přístup využívají některé algoritmy, které samy o sobě obsahují výběr atributů jako součást procesu učení. Mezi ně patří rozhodovací stromy nebo SVM (metoda podpůrných vektorů). [46][56][57]

Kromě výběru nejrelevantnějších atributů a odstranění ostatních je možné snižovat příznaky v souboru pomocí redukce dimenzionality. Tato technika snižuje rozměrnost dat, ale snaží se o zachování co nejvíce informací. Výsledkem redukce dimenzionality je dataset s menším počtem atributů, které reprezentují lineární kombinace původních dat, případně je zastupují jinak (jsou tedy vytvořeny nové atributy, které nahrazují větší počet původních příznaků). [56][57]

3.3.5 Rozhodovací stromy

Ve strojovém učení patří rozhodovací stromy (decision trees) mezi nejpoužívanější algoritmy. Jsou složeny z uzlů a hran, které představují grafickou reprezentaci sady pravidel pro klasifikaci. Uzly v tomto orientovaném grafu reprezentují testy určitého atributu datové sady (podmínku) a každá z hran výsledky těchto testů. Při průchodu sestaveným grafem od kořene lze pro jakýkoli validní vstup pomocí vyhodnocování podmínek na uzlech vybrat jeden list, který reprezentuje konkrétní třídu. Proces učení nad trénovacím datasetem v tomto případě znamená sestavování stromu na základě vzorců a vztahů mezi daty v souboru. [46][58][59]

Jednoduchý algoritmus pro vytváření stromu se může skládat z následujících kroků:

1. výběr atributu, který nejlépe rozděluje danou sadu dat na podmnožiny (podle míry relevance při rozhodování o příslušnosti ke třídě);
2. vytvoření uzlu, který testuje hodnotu vybraného atributu;
3. rozdělení datasetu do podmnožin na základě hodnoty vybraného atributu;
4. opakování procesu s každou z vytvořených podmnožin, dokud nebyl dosažen maximální počet uzlů nebo nepatří-li všechny prvky jedné třídě. [58][59]

Existuje mnoho různých algoritmů, které jsou využívány pro vytváření těchto rozhodovacích stromů. Liší se často tím, jestli vytváří binární či vícecestné stromy. Odlišný je také způsob vyhodnocování relevance atributu na rozhodování o příslušnosti vstupu ke třídě. Aby se předešlo overfittingu, algoritmy často pracují s předem definovanými

limity jako je například omezení hloubky stromu nebo minimální počet prvků na koncovém listu. [45][46][58]

Výhodou rozhodovacích stromů je snadná interpretovatelnost, což umožňuje snadnou vizualizaci a porozumění predikcím modelu. Kromě samotné předpovědi jsou tedy analyzovatelná i jednotlivá rozhodnutí, která k ní vedla. Tyto algoritmy jsou zároveň efektivní při učení (trénování) a zejména při predikci, což umožňuje rychlé zpracování velkých objemů dat. [46][59]

Mezi hlavní nevýhody patří sklony k přílišnému přizpůsobení modelu trénovacím datům (a špatné výkonnosti na testovacích datech) a nestabilita – i malé změny v testovacích datech mohou vést k velmi odlišným stromům. Při nevhodném nastavení pro konkrétní klasifikační problém také mohou mít stromy omezenou přesnost. [46][59]

3.3.6 Náhodný les

Náhodný les (random forest) je často využívaná metoda strojového učení, která ke klasifikaci využívá více rozhodovacích stromů. Každý z těchto stromů je sestaven nad podmnožinou dat, která je vytvořena technikou bootstrappingu (z trénovacích dat se vybírá náhodně s opakováním). Tyto nové soubory pro učení jednotlivých stromů jsou stejně velké, jako původní dataset. Některé implementace algoritmu náhodného lesa také využívají náhodný výběr z atributů. Při vytváření predikcí pro vzorek je využíván princip hlasování. Třída je určena každým rozhodovacím stromem zvlášť, přičemž nejčastěji zastoupená kategorie se stane výsledkem klasifikace. [46][60]

Před vytvářením modelu pomocí náhodného lesa je definován hyperparametr, který definuje, kolik bude vytvořeno rozhodovacích stromů. Příliš nízký počet by znamenal nekvalitní predikce, ale čím více stromů model využívá, tím je jeho sestavení výpočetně náročnější. Zároveň je možné definovat parametry pro jednotlivé stromy (např. maximální hloubka nebo minimální počet prvků na listu). [46][60][61]

Výhodou náhodného lesa je vysoká přesnost predikcí díky kombinaci více rozhodovacích stromů. Tato architektura determinuje vysokou robustnost a odolnost vůči šumu nebo odlehlým hodnotám. Algoritmus je velmi rychlý a umožňuje efektivně zpracovávat velké datové sady. Stejně tak jako rozhodovací stromy může pracovat i s kategorickými

proměnnými. Zároveň je relativně odolný k přílišnému přizpůsobení modelu trénovacím datům. [60][62]

Nevýhody této metody spočívají v menší interpretovatelnosti (kombinuje mnoho jednotlivých stromů) a dále také ve velké závislosti na předem definovaných parametrech. Ty mohou zásadním způsobem ovlivnit jeho výkon – volba správných parametrů je tedy klíčová. [60][62]

3.3.7 Naive Bayes

Dalším algoritmem, který lze pro klasifikaci pomocí strojového učení využít, je Naive Bayes, který předpovídá pravděpodobnost, že daný vstupní objekt patří do určité třídy na základě kombinace vstupních atributů. Algoritmus předpokládá, že mezi žádnými dvěma proměnnými v souboru neexistuje závislost. Během procesu učení je pro každý atribut vypočítáno, jak jeho hodnoty ovlivňují příslušnost ke konkrétní třídě (je vypočítána pravděpodobnost podmíněného jevu A za podmínky jevu B pomocí Bayesova pravidla). To znamená, že pravděpodobnost příslušnosti celého vstupu k určité kategorii lze získat výpočtem za pomoci jednotlivých pravděpodobností pro každý z atributů. [46][63][65]

Naive Bayes lze tedy využít k binární i vícetřídní klasifikaci. Princip jeho fungování umožňuje snadnou klasifikaci i takových vstupních objektů, ve kterých některé hodnoty chybí. Pokud jsou vstupní atributy numerické a spojité, je vhodné použít Gaussovský Naive Bayes (pravděpodobnosti pro jednotlivé vstupní atributy jsou modelovány pomocí normálního rozdělení). Existuje i mnoho jiných implementací algoritmu Naive Bayes, které se liší zejména v požadavcích na typy atributů ve vstupním souboru. [46][63][65]

Algoritmus je typický výhodami jako jednoduchost na implementaci a použití. Platí, že je efektivní i v rozsáhlých datových sadách a také při velkém počtu atributů. Pro některé typy úloh má zároveň velmi dobré výsledky. [46][64]

Hlavní nevýhodou pro určité klasifikační problémy je předpoklad nezávislosti jednotlivých proměnných, což může vést k chybám a nízké výkonnosti v případě, že mezi atributy existují tyto závislosti. Naive Bayes tedy není vhodným řešením, pokud je od

výsledného modelu očekáváno, že bude zohledňovat složité vztahy mezi proměnnými. [46][64]

3.3.8 Algoritmy podpůrných vektorů

Algoritmy podpůrných vektorů (Support Vector Machines – SVM) patří mezi nejrozšířenější metody strojového učení. Jsou založeny na hledání hyperroviny, která co nejlépe odděluje jednotlivé třídy dat v mnohorozměrném prostoru. [46][66][67]

Proces učení nad trénovacími daty zahrnuje transformaci do vyššího rozměru (z důvodu snazší separovatelnosti). Následně jsou pro každou třídu hledány podpůrné vektory patřící do různých tříd, přičemž se hledá se optimální hyperrovina, která odděluje tyto vektory. Ta je nalezena tak, že se minimalizuje chyba klasifikace na trénovacích datech a zároveň se maximalizuje vzdálenost mezi podpůrnými vektory a hyperrovinou. Výsledná klasifikace proběhne podle toho, na které straně hyperroviny se každý z testovaných objektů nachází. [46][66][67]

Mezi výhody SVM patří poměrně dobrá klasifikační schopnost, odolnost proti šumu, rychlost (efektivita) učení i nad velkými datovými soubory a flexibilita, která umožňuje volbu vlastní jádrové funkce využívané pro transformaci dat do vyššího rozměru. [66][67]

Nevýhodou těchto algoritmů je (časová) náročnost na výběr parametrů, která někdy vede k tomu, že se ideální nastavení vůbec nepodaří najít. Dále jsou tyto metody typické špatnou interpretovatelností výsledků (zvláště souborů s mnoha atributy prakticky nelze vysvětlit, na základě jakých rysů bylo o výsledné třídě rozhodnuto). Při používání SVM je třeba pracovat s vyváženými datasety, protože jsou citlivé na nerovnováhu tříd. [66][67]

3.3.9 Umělé neuronové sítě

Neuronové sítě představují v kontextu umělé inteligence modely, které jsou inspirovány biologickými strukturami, oproti kterým jsou většinou poměrně zjednodušené. V současné době jsou tyto struktury nejčastěji sestavovány za účelem řešení konkrétní a úzce vymezené úlohy. Neuronové sítě patří mezi algoritmy strojového učení – jejich cílem

je tedy osvojit si vzorce a vztahy přítomné v trénovacích datech a na základě nich vyvozovat z nových dat závěry nebo předpovědi. [47][68][69]

Základní stavební jednotkou umělých neuronových sítí je neuron, který přijímá a zpracovává informace. Každý neuron se skládá z následujících částí:

- jeden nebo více vstupů (dendrity),
- váhy na vstupech,
- aktivační funkce,
- výstup neuronu (axon). [47][68][69]

Na vstupech neuronu jsou signály, které přicházejí z jiných neuronů nebo z vnějšího prostředí. Každý vstup má vlastní váhu, která reprezentuje míru důležitosti tohoto konkrétního vstupu. Mohou být kladné i záporné, což v důsledku ovlivňuje, zda daný vstup podporuje aktivaci nebo inhibici celého neuronu (v kontextu aktivační funkce). Váhy jsou před začátkem učení většinou stanoveny náhodně a během trénování sítě jsou aktualizovány. [47][68][69]

Poté je stanovena vážená suma signálů (skalární součin vektoru vah s vektorem hodnot na vstupech). Výsledný skalární součin může být ještě rozšířen o tzv. práh (threshold), jehož hodnota se k součinu přičte. V některých případech je práh konstantní a stanovený před začátkem učení, ale v současných implementacích neuronových sítí je často jeho hodnota průběžně aktualizována během trénování. Tento výpočet je vyjádřen rovnicí 6, kde x_i je vstup, w_i váha a θ hodnota prahu. [47][69][70]

$$Z = \sum_{i=1}^{n} x_i w_i + \theta \tag{6}$$

Výsledná hodnota Z získaná pomocí rovnice 6 představuje vstup pro aktivační funkci, která určuje zda a jak silně bude neuron aktivován. Existuje více typů těchto funkcí a konkrétní výběr závisí na poloze neuronu v neuronové síti a na řešeném problému. Vypočítaná funkční hodnota je použita jako výstup neuronu. Mezi nejběžnější aktivační funkce patří:

- sigmoidní funkce – má tvar křivky, která se na ose y pohybuje od 0 do 1 a je vhodná pro binární klasifikaci, kde jsou výstupní hodnoty omezeny na toto rozmezí,
- ReLU (rectified linear unit) – vrací 0 pro záporný vstup a vstupní hodnotu beze změny pro všechny kladné vstupy; je v současnosti často využívaná a zefektivňuje proces učení,
- tanh (hyperbolický tangens) – tvar křivky je podobný sigmoidní funkci a je využívána zejména pro regresní úlohy,
- softmax – ve vícetřídních klasifikačních problémech vhodná pro výstupní neurony, kde pokud každý zastupuje jednu třídu, vrátí funkce na každém z nich pravděpodobnost příslušnosti k této třídě (hodnota mezi 0 až 1). [68][69][70]

Proces učení v neuronových sítích znamená minimalizaci hodnoty ztrátové (chybové) funkce, která reprezentuje celkovou chybu vyhodnocenou na základě rozdílů mezi predikovanými hodnotami a skutečnými hodnotami v trénovacích datech. Chybová funkce je volena podle toho, o jaký problém se jedná. Mezi nejčastěji používané patří střední kvadratická chyba (Mean Squared Error – MSE) a při řešení klasifikačních problémů lze uplatnit křížovou entropii (Cross Entropy). [70][71][72]

Jakmile je známa průběžná hodnota chybové funkce, jsou upraveny váhy na vstupech jednotlivých neuronů v síti. To je obvykle (ve většině implementací) realizováno pomocí algoritmu zpětného šíření chyby (backpropagation). Postupuje se od výstupních neuronů směrem ke vstupu a u vah jednotlivých vstupů neuronů je zjišťován podíl na celkové chybě sítě (jak moc změna konkrétní váhy ovlivňuje výstup sítě). K tomu se využívá metoda gradientního sestupu, která je schopná pomocí derivace ztrátové funkce určit směr, kterým je třeba váhu upravit, aby byla minimalizována chyba. Nevýhodou tohoto přístupu je riziko stagnace v lokálním minimu a z toho plynoucí nesnižování chyby sítě. [47][70][71]

Jednotlivé neurony jsou v rámci neuronové sítě organizovány ve vzájemně propojených vrstvách:

1. vstupní vrstva – na vstupy jejích neuronů přichází vstupní data pro síť (hodnoty atributů z datasetu),

2. skryté vrstvy – jsou umístěny mezi vstupní a výstupní vrstvou a jejich hlavním účelem je extrahovat potřebné informace ze vstupů,
3. výstupní vrstva – generuje výstupní hodnoty. [69][72]

Každá z vrstev může obsahovat různý počet neuronů v závislosti na řešené úloze. Typické je, že mají neurony v rámci jedné vrstvy stejnou aktivační funkci. Počet vrstev, jejich vzájemné propojení a další parametry je vhodné zvolit podle toho, jaká je komplexnost a druh řešeného problému, přičemž sítě s vyšším počtem skrytých vrstev jsou označovány jako „hluboké“. [70][71][72]

Nalezení vhodných parametrů a počáteční konfigurace neuronových sítí jsou obtížné a liší se s každou úlohou. Ve většině případů neexistuje předem známé optimální řešení, které by bylo možné implementovat. K návrhu architektury některých částí neuronové sítě lze ale využít již existující a ověřené vzory. Jak již bylo zmíněno, například pro vícetřídní klasifikaci je vhodné mít ve výstupní vrstvě neurony zastupující každou z tříd, přičemž při využití aktivační funkce softmax je možné získat pravděpodobnosti příslušnosti k jednotlivým třídám. Zároveň je možné vhodné počáteční parametry určit pomocí různých aproximačních technik, mezi které se řadí mimo jiné evoluční algoritmy. [47]

3.3.10 Evoluční algoritmy

V kontextu umělé inteligence a strojového učení se evoluční algoritmy uplatňují v řešení optimalizačních úloh, přičemž se inspirují v biologických systémech. Vytvářejí populaci jedinců, kteří představují různá potenciální řešení zvoleného problému. Z těchto jedinců jsou vybíráni nejlepší (podle schopností řešit danou úlohu) a pomocí technik, jako je například křížení a mutace, je z nich vytvořena nová populace. Tento proces se opakuje a předpokládá se, že nově vzniklí jedinci budou mít v průměru lepší vlastnosti, než předchozí generace. [47][72]

Typickou optimalizační úlohou je hledání počátečního nastavení algoritmu strojového učení. Ve většině případů je z důvodu výpočetní náročnosti prakticky nemožné vyzkoušet všechny kombinace parametrů a nalézt exaktní řešení. Často je tedy vhodným přístupem při návrhu počátečního nastavení využít spojení již otestovaných a funkčních

vzorů s evolučními algoritmy. Není zaručeno, že tyto aproximační techniky naleznou optimální řešení úlohy, ale je pomocí nich možné v relativně krátkém čase získat takové řešení, které se optimu blíží. [47][72]

Jeden z nejčastěji používaných algoritmů je genetický evoluční algoritmus (GEA). Pracuje s populací jedinců, tedy možných řešení úlohy, kteří jsou reprezentováni jako chromozomy složené z genů. Počáteční populaci lze vygenerovat náhodně, ale může obsahovat i předem jednoznačně definované jedince. Je využita tzv. fitness funkce, která pro každého jedince jednoznačně ohodnocuje jeho kvality v kontextu řešení zadaného problému a potažmo tedy určuje schopnosti přežít v okolním prostředí. Dalším krokem je selekce k následnému křížení, přičemž platí, že čím lepší je jedinec (podle fitness funkce), tím vyšší má šanci být vybrán. [47][72]

Proces křížení je vytváření nových jedinců na základě kombinace chromozomů rodičů. Zároveň je možné provést náhodnou změnu (mutaci) určité části nového genomu. Algoritmus křížení se liší podle reprezentace jednotlivých genů. V případě binárního genomu se používá tzv. maska, která určí, které části bude mít nový jedinec od každého z rodičů. Geny reprezentované reálnými čísly je možné kombinovat například pomocí určení hodnoty v intervalu mezi oběma rodiči. GEA má mnoho variant implementace v závislosti na typu řešené úlohy. [47][72]

Další skupina evolučních algoritmů využívá tzv. inteligenci hejna. Vychází z myšlenky, že přestože jednotlivé částice interagují pouze lokálně se svým okolím a mezi sebou a neexistuje žádná centralizovaná struktura řízení, celý systém vykazuje složité globální chování. Tyto algoritmy většinou zahrnují částice pohybující se v prostoru podle jednoduchých pravidel, přičemž poloha každé z nich odpovídá konkrétnímu řešení dané úlohy ohodnocenému pomocí fitness funkce. Existuje mnoho různých implementací, které jsou často inspirovány systémy, jako například mravenčí kolonie, včelí roj nebo hejna ryb či ptáků. Jednotlivé algoritmy vycházející z principů inteligence hejna mohou či nemusí využívat principy křížení. [72][73][74]

4 Vlastní práce

Přestože užívání psychedelických látek téměř nemá fyzické dopady na jejich uživatele, stále zde existují určitá rizika spojená s psychikou. Z krátkodobého hlediska se jedná zejména o nepříjemný zážitek. Pokud jde o velmi náročnou zkušenost, někteří lidé mohou mít i po jejím odeznění určité problémy s její integrací. Jak již bylo zmíněno, psychedelické látky jsou velmi specifické tím, že na průběh zkušenosti má podstatný vliv set a setting, tedy kombinace psychického stavu jedince (předchozí zkušenosti, představy nebo různá očekávání) a prostředí, kde se během zkušenosti nachází. Z toho důvodu nelze stanovit všeobecně platná pravidla užívání, která by zaručovala, že zkušenost proběhne dobře a bez jakýchkoli rizik.

Charakter psychedelické zkušenosti lze do určité míry zpětně kvantifikovat pomocí dotazníku. Díky tomu je možné měřit a porovnávat jednotlivé aspekty průběhu zážitku – jak negativní, tak i pozitivní. V dostatečně kvalitních záznamech o různých psychedelických zkušenostech by měla tedy existovat určitá korelace mezi průběhem zážitku a počátečním setem, settingem a dalšími informacemi o uživateli. Pokud by byl tento problém formulován jako klasifikační nebo regresní úloha, mělo by být možné ji řešit pomocí metod strojového učení.

Cílem praktické části práce je tedy vytvořit predikční model, který potenciálním uživatelům psychedelických látek ještě před samotným zážitkem s určitou přesností poskytne informace o charakteru případné psychedelické zkušenosti. Tito lidé budou tedy předem znát pravděpodobnou míru rizika, tedy predikci pozitivních i negativních aspektů zážitku a budou moci lépe zhodnotit, zda si chtějí touto zkušeností projít.

V rámci teoretických východisek bylo zmíněno, že psychedelická zkušenost je poměrně špatně predikovatelná. Hlavním důvodem jsou nedostatečné možnosti pro exaktní popis a kvantifikaci setu, tedy psychického stavu uživatele. Zároveň je možné předpokládat, že data získaná pomocí dotazníků budou obsahovat mnoho nepřesností a šumu – ať už z důvodu neúmyslných chyb při vyplňování, tak i kvůli záměrnému nepřiznání některých faktů nebo naopak jejich nadhodnocování. Pravděpodobnou špatnou predikovatelnost způsobenou těmito nedostatky bude nutné v práci zohlednit.

4.1 Zdroj dat

Data, pomocí nichž budou vytvářeny predikční modely, byla sesbírána pomocí dotazníkového šetření za využití mobilní aplikace iTrip, projektu Nadačního fondu pro výzkum psychedelik (PSYRES). Aplikace byla navržena za účelem výzkumu psychedelických látek na základě hodnocení jednotlivých zkušeností a jejich kontextu. Zároveň svým uživatelům poskytuje informace o různých psychedelikách a jejich účincích. Cílí na širokou populaci lidí, od příležitostných uživatelů až po pacienty, kteří využívají tyto látky pravidelně pro svou léčbu. Vzhledem k ochraně osobních údajů lze zveřejňovat pouze výstupy po statistickém zpracování. [75] Autor této práce se na vývoji aplikace iTrip a vytváření dotazníků nepodílel.

Záznamy o psychedelických zkušenostech, které budou analyzovány, byly sbírány průběžně od března 2020 do listopadu 2022. Tato diplomová práce je v době svého vzniku jediná, která využívá a zpracovává data z aplikace iTrip. Neexistují proto žádné zdroje, které by mohly být použity k získání informací o povaze těchto dat. Výsledky a výkonnost predikčních modelů tedy nebude možné porovnávat s jinými pracemi.

4.2 Zpracování dat

Ke zpracování dat je využíván primárně Python, v němž jsou dostupné knihovny jako NumPy nebo Pandas, které slouží k rychlé a efektivní práci s daty. Tato volba je výhodná také z důvodu, že vytváření predikčních modelů bude probíhat ve stejném prostředí a odpadá tak nutnost připravený výstup exportovat a přenášet jinam. Dalším zvoleným nástrojem je MySQL, které uplatňuje relační databázový model, protože jazyk SQL je efektivním prostředkem pro provádění některých typů operací nad daty.

Aplikace iTrip využívá objektovou databázi, která je hostována na Google Firebase. Vyexportovaná data jsou tedy strukturována ve formátu JSON a jde celkem o tři soubory. Vzhledem ke snadné zpracovatelnosti v jazyce JavaScript se jedná o praktickou volbu pro REST API pro mobilní či webové aplikace. Tento formát umožňuje reprezentovat strukturovaná data jako jsou objekty a pole a zároveň podporuje zanořování do neomezené úrovně. Pro algoritmy strojového učení je ale vhodná spíše reprezentace

pomocí dvourozměrné matice.

Prvním krokem je tedy převedení dat uložených jako JSON do formátu CSV, který je určen pro tabulkovou (dvourozměrnou) reprezentaci. Aplikace byla za dobu své existence několikrát aktualizována, což způsobilo určité nekonzistence v datech (různé formáty v rámci jedné proměnné), které je potřeba před další prací odstranit. Vhodným řešením pro tuto operaci je Python. Po převedení původních třech souborů do CSV jsou nahrány jako tři tabulky do relační databáze (MySQL) k dalšímu zpracování.

4.2.1 Tabulky

Z důvodu přehlednosti tato práce neuvádí seznam a popis všech atributů před začátkem zpracování – tabulky mají 19, 259 a 179 sloupců. V rámci této kapitoly budou tedy nejprve stručně popsány tyto jednotlivé datasety a úpravy, které je nutné provést před použitím různých algoritmů strojového učení. Výčet atributů a jejich význam bude uveden až pro zpracovaný a připravený soubor.

První z těchto tabulek – **users** – obsahuje 19 atributů a 2092 instancí. Jde o údaje o uživatelích jako je vzdělání, finanční situace, víra, pohlaví, věk nebo výška a váha. Každý uživatel má unikátní kód, který zde funguje jako primární klíč. Záznamy v ostatních tabulkách mohou tedy nést referenci na konkrétního člověka pomocí této hodnoty.

Důležité údaje jsou také v tabulce **substances**. Obsahuje záznamy o osobnostních rysech, psychickém stavu člověka i o případné předchozí psychologické či psychiatrické pomoci. Zároveň nese informace o osobní historii užívání různých látek (kromě psychedelik také stimulanty, opiáty, alkohol, konopí nebo tabák). Tento dataset taktéž mapuje závislosti. Míra návyku je pro každou z látek zjišťována pomocí odpovědí na následující otázky:

1. Jak často jste užíval/a látku XY během posledních 12 měsíců?
2. Cítil/a jste někdy potřebu užívání látky XY snížit?
3. Užíval/a jste současně další psychoaktivní látku s látkou XY?
4. Jste vždy když chcete schopen/schopna přestat s užíváním látky XY?
5. Měl/a jste „okna“ nebo „flashbacky“ v důsledku užívání látky XY?

6. Měl/a jste někdy výčitky nebo pocity viny kvůli užívání látky XY?
7. Vyčítali vám někdy partner nebo rodiče vaše užívání látky XY?
8. Zanedbával/a jste někdy svou rodinu kvůli užívání látky XY?
9. Dopustil/a jste se někdy nelegálního jednání abyste získal/a látku XY?
10. Měl/a jste někdy abstinenční příznaky když jste přestal/a brát látku XY?
11. Užil/a jste někdy látku XY ihned po ránu abyste se uklidnil/a nebo se zbavil/a kocoviny nebo abstinenčních příznaků?
12. Měl/a jste zdravotní problémy v důsledku užívání látky XY?

Protože se v aplikaci zobrazují pouze relevantní otázky na základě předchozích odpovědí, v tabulce substances chybí mnoho hodnot. Vzhledem k tomu, že je zjišťována závislost na všech látkách odděleně, v datasetu je 259 atributů. Záznamů je 1744; ne všichni uživatelé dotazník vyplnili a zároveň je možné jej projít vícekrát (v dalším časovém horizontu). Tabulka také obsahuje referenci na uživatele (cizí klíč do users).

Jednotlivé zkušenosti jsou zaznamenány v tabulce **experiences**. Konkrétně jde o látku a způsob jejího užití, motivace k podstoupení zkušenosti nebo také setting (místo užití, lidé v okolí, hudba). Dataset zároveň obsahuje evaluaci zkušenosti z hlediska pětidimenzionální stupnice hodnocení změněných stavů vědomí, míry rozpuštění (a zbytnění) ega a také mystických aspektů. V tabulce jsou také základní informace týkající se reflexe zážitku a ovlivnění dalšího života. Vzhledem k rozsáhlým dotazníkům je zde 179 atributů a 806 instancí. I zde platí, že ne všichni uživatelé dotazník vyplňovali a někteří ho naopak mohli vyplnit vícekrát. Stejně jako v substances je přítomna reference do tabulky users (cizí klíč).

4.2.2 Úpravy dat v tabulkách z hlediska řádků (záznamů)

Z hlediska řádků je hlavní úpravou odstranění problematických instancí. Prvním krokem je eliminace záznamů, kde chybí velké množství hodnot. Pro naprostou většinu instancí v těchto tabulkách platí, že jsou buď téměř všechny hodnoty známé, nebo naopak téměř všechny chybí. Rozpoznání řádků vhodných k odstranění tedy není komplikované a jsou smazány ty, kde je vyplněno méně než 50 %.

V aplikaci byla mezi látkami uvedena fiktivní droga relevin. Záznamy, které její užívání zmiňují, mohou být také odstraněny. Lze totiž předpokládat, že lidé, kteří v dotaznících uvedli zkušenost s relevinem, nevyplňovali formuláře pravdivě. Smazány byly taktéž instance vytvořené vývojáři aplikace za testovacími účely. Je velmi pravděpodobné, že tato data (vzhledem ke způsobu jejich sběru) obsahují mnoho nenalezených nepřesností či nepravd, ale jejich eliminace je prakticky neproveditelná z důvodu nemožnosti identifikace.

Počty instancí v jednotlivých tabulkách jsou prezentovány v tabulce 4, která ukazuje stav před zpracováním (čištěním) a po něm.

Tabulka	Před zpracováním	Po zpracování
users	2092	2082
substances	1744	1713
experiences	806	704

Tabulka 4: Počty instancí v tabulkách

Zdroj: vlastní zpracování

Před použitím datasetu pro sestavení predikčního modelu je důležité ověřit, zda je vyvážený – v případě klasifikace to znamená, že ke každé ze tříd náleží podobný počet instancí. V této fázi zpracování v datovém souboru ještě nejsou vytvořeny atributy, jejichž hodnota bude predikována. Autor práce se proto zabývá problematikou vyváženosti v dalších kapitolách.

4.2.3 Úpravy dat v tabulkách z hlediska sloupců (atributů)

Nejprve je nutné odstranit sloupce, ve kterých příliš mnoho hodnot chybí. V případě ponechání by pravděpodobně měly negativní vliv na efektivitu učení a výkonnost vytvořených modelů. Tento krok není ale možné provést čistě automaticky pro všechny atributy, protože:

1. v některých sloupcích může být nevyplněná hodnota relevantní (např. nevyplněné údaje o hudbě pravděpodobně znamenají žádnou hudbu),

2. určité sloupce budou seskupovány (např. ukazatele závislosti), takže je odstranění žádoucí pouze tehdy, když není vyplněna žádná hodnota v rámci této skupiny.

Kromě sloupců s chybějícími hodnotami jsou odstraněny také ty, které mají nevhodný datový typ pro další práci. Konkrétně jde o čistě textové atributy, kterými v tomto datasetu jsou například volný slovní popis zkušenosti nebo žánr hudby poslouchané během zážitku.

Aby bylo možné data později použít při práci s různými algoritmy strojového učení, je důležité, aby byly všechny atributy numerické. Hodnoty v kategorických textových proměnných, které jsou zároveň ordinální (mezi hodnotami existuje pořadí), je možné převést na numerické pomocí prostého nahrazení číselnou řadou, která toto pořadí respektuje. Pokud někde hodnota chybí, podle lze podle situace postupovat dvěma způsoby:

1. pokud v kontextu atributu dává prázdná hodnota smysl a má místo v pořadí hodnot, je možné ji nahradit číslem s ohledem na tuto řadu,
2. v opačném případě lze na místo chybějící hodnoty vložit většinově zastoupenou kategorii.

U ostatních kategorických proměnných, mezi jejichž hodnotami nelze pořadí nalézt, je zapotřebí rozdělení do více sloupců. To je provedeno tak, že je vytvořen nový atribut pro každou možnou kategorii. Tyto nové sloupce poté obsahují v jednotlivých instancích hodnoty 0 a 1 v závislosti na tom, jaká byla původní kategorie. Tento princip je detailněji popsán v rámci teoretických východisek práce a je použitelný i v situacích, kdy je jedna hodnota reálně složena z více položek (otázka, ke které lze v dotazníku vybrat více odpovědí).

Také u numerických proměnných je nutné vyřešit problém chybějících hodnot. Nejlepší volbou je vyplnění těchto prázdných míst mediánem vypočítaným z konkrétní proměnné, který na rozdíl od průměru nezohledňuje extrémy. Je vhodné zachovat datový typ – pokud sloupec obsahuje pouze celočíselné hodnoty a medián je desetinné číslo, bude zaokrouhlen.

Kromě toho bývá u numerických proměnných často řešena problematika rozsahů hodnot. Jak již bylo zmíněno v teoretické části práce, pro většinu algoritmů je problematické, když se v tomto ohledu příznaky výrazně liší. Proto jsou všechny atributy normalizovány tak, aby jejich hodnoty vždy ležely mezi 0 a 1 (s ohledem na zachování poměrově stejných vzdáleností mezi nimi).

Žádoucí také je provést určitou redukci dimenzionality. Některé skupiny sloupců obsahují informace o jednom jevu a jejich sloučení má (kromě zpřehlednění) pozitivní vliv na efektivitu při zpracování pomocí algoritmů strojového učení. K této operaci je ale nutné přistupovat individuálně pro každou takovou množinu atributů. Tyto skupiny jsou od sebe navzájem odlišné a sloučení tedy není možné provést vždy stejným způsobem. Sdružení sloupců je mimo jiné vhodné například v tabulce substances, konkrétně u ukazatelů závislosti. Jak již bylo zmíněno, míra návyku je měřena pomocí zjišťovacích otázek. Za předpokladu, že je „ano“ a „ne“ nahrazeno pomocí 1, respektive 0, může být vytvořen nový sloupec, který bude obsahovat součet těchto hodnot napříč atributy pro každou zkoumanou látku zvlášť.

Na výkonnost modelů a rychlost učení má podstatný vliv počet příznaků v souboru. Tudíž je vhodné odstranit ty, které mají (při klasifikaci) velmi malý (nebo žádný) vliv na přiřazení instance ke třídě. Tato práce se problematikou výběru atributů zabývá v dalších kapitolách.

Po úpravách došlo ke změnám v počtu atributů, jak ukazuje tabulka 5. Výrazný rozdíl v substances je způsobený zejména slučováním u atributů vyjadřujících různé znaky závislosti. Velký počet příznaků v experiences nepředstavuje problém, protože většina sloupců popisuje průběh zkušenosti, což bude předmětem dalšího zpracování.

Tabulka	Před zpracováním	Po zpracování
users	19	17
substances	259	54
experiences	179	198

Tabulka 5: Počty atributů v tabulkách

Zdroj: vlastní zpracování

4.3 Formulace predikční úlohy

Aby bylo možné předvídat charakter psychedelické zkušenosti pomocí metod strojového učení, je nutné tento problém zformulovat jako predikční úlohu. Na základě dostupných dat je možné proběhlé zážitky umístit na pětidimenzionální stupnici hodnocení změněných stavů vědomí (5D-ASC) a kromě toho je lze hodnotit i z hlediska míry rozpuštění ega. Přestože jsou v tabulce experiences také příznaky popisující mystické aspekty zážitků, jsou jejich hodnoty známy v méně než polovině instancí, což je pro strojové učení příliš nízký počet.

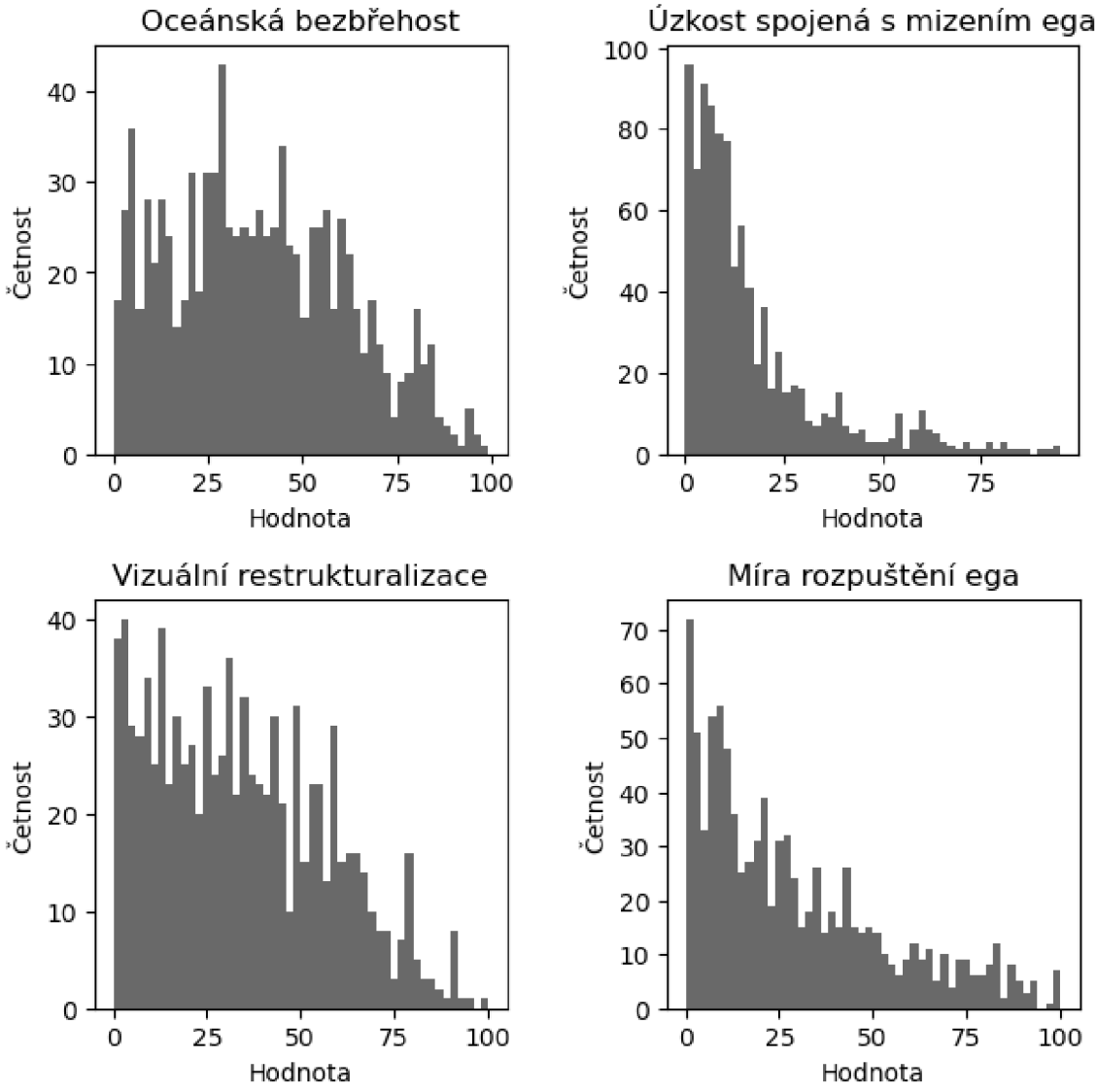
Z hlediska snižování rizik při užívání psychedelik je vhodné předem znát pravděpodobné zařazení zkušenosti do prvních třech dimenzí 5D-ASC: oceánská bezbřehost (OBN), úzkost spojená s mizením ega (DED) a vizuální restrukturalizace (VRS). Ty je možné doplnit o míru rozpuštění ega (EDI), která sama o sobě není pozitivním ani negativním aspektem, ale hodnotí spíše intenzitu zkušenosti. Jak bylo popsáno v rámci teoretických východisek práce, konkrétní hodnoty těchto čtyřech ukazatelů je možné získat pomocí ohodnocení příslušných tvrzení, což je v aplikaci implementováno jako škály s rozmezím od 0 do 100. Z těchto údajů lze pro jednotlivé instance spočítat aritmetický průměr reprezentující každou ze zmíněných metrik.

Na základě těchto úvah je zřejmé, že model bude predikovat toto zařazení na základě příznaků, jejichž hodnoty jsou známy ještě před samotným započetím zkušenosti. Problém by bylo teoreticky možné řešit jako klasifikační úlohu, ale i pomocí regrese. Při rozhodování mezi těmito dvěma variantami je nutné zohlednit špatnou predikovatelnost vycházející jak z povahy dat, tak z poměrně nízkého počtu instancí. Proto se klasifikace v této situaci jeví jako vhodnější volba.

4.3.1 Vytvoření tříd pro klasifikaci

Rozdělení pravděpodobnosti závislých proměnných je velmi nerovnoměrné. Z histogramů (obrázek 1) je zřejmé, že se většina hodnot nachází v levých částech grafů. To je možné vysvětlit tím, že jsou zážitky různorodé, přičemž stejné skóre v určité dimenzi může vycházet z odlišného ohodnocení různých tvrzení. Zkušenosti s podobným zařazením jsou z obecného hlediska kvalitativně srovnatelné, ale jejich reálný průběh

nemusí být totožný. Dále také platí, že se v průběhu mnohdy střídají různé stavy – pokud se například objeví nepříjemné pocity (úzkost spojená s mizením ega), typicky po určité době odezní. Při zpětném hodnocení zážitku pomocí dotazníku pak nedává smysl je označovat na škále vysoko, pokud se objevily pouze na krátkou dobu. To má za následek velké množství nízkých hodnot DED.



Obrázek 1: Histogramy distribuce hodnot proměnných

Zdroj: vlastní zpracování

Z toho vyplývá, že zkušenosti nelze uspokojivě hodnotit jednotlivě pomocí absolutně vyjádřených hodnot. Zařazení zážitku v dimenzích má význam zejména z hlediska relativního porovnávání. Proto by vytvoření tříd pomocí prostého rozdělení intervalu $\langle 0, 100 \rangle$ na stejně velké části nebylo v tomto případě příliš smysluplné. Mnohem vhod-

nější je použít kvantily, které ze své definice obecně respektují distribuci hodnot v rámci proměnných. Konkrétní predikce příslušnosti ke kvantilu bude tedy ukazovat pravděpodobné zařazení mezi ostatními zkušenostmi a nikoli absolutní údaje. Velkou výhodou tohoto přístupu je také zajištění vyváženosti datasetu, protože instance budou rovnoměrně rozloženy mezi třídami.

Než bude možné vytvořit třídy, je nutné určit, jaké kvantily budou použity. Vzhledem k předpokládané špatné predikovatelnosti je vhodné zvolit spíše nižší počet kategorií. Nejvyšší výkonnosti modelů by mělo být teoreticky dosaženo při použití mediánu a tedy rozdělení do dvou tříd. Tato volba by však znamenala poměrně nízkou informační hodnotu předpovědí. Vhodnější jsou proto tercily – rozdělení na tři části. Mezi těmito kategoriemi lze navíc stanovit pořadí, což znamená, že vícetřídní klasifikace v tomto případě nabízí kromě standardních výkonnostních metrik použití rozšířené přesnosti.

Hodnoty tercilů pro jednotlivé závislé proměnné jsou uvedeny v tabulce 6. Stejně jako z histogramů je z těchto indikátorů zřejmé velmi nerovnoměrné rozdělení souboru. V datasetu jsou tedy nahrazeny původní numerické hodnoty třídami *low*, *mid* a *high* podle příslušnosti k tercilu.

Proměnná	$Q_{1/3}$ (první tercil)	$Q_{2/3}$ (druhý tercil)
Oceánská bezbřehost (OBN)	26	48
Úzkost spojená s mizením ega (DED)	7	15
Vizuální restrukturalizace (VRS)	20	43
Míra rozpuštění ega (EDI)	11	25

Tabulka 6: Hodnoty tercilů pro proměnné

Zdroj: vlastní zpracování

4.4 Výsledný dataset

Po zpracování dat v tabulkách, formulaci klasifikačního problému a navržení jednotlivých tříd je možné data použít pro sestavování modelů strojového učení. Ke spojení tabulek je možné použít operaci INNER JOIN, protože substances a experiences obsahují

referenci do tabulky users. Zároveň je náhodně změněno pořadí instancí a skryty atributy nevýznamné z hlediska další práce (např. primární klíče původních tabulek nebo datum registrace uživatele). Výsledný dataset má celkem 933 instancí a obsahuje 110 atributů. První skupina příznaků se týká obecných informací a osobní historie:

1. **education** – nejvyšší dosažené vzdělání (1 až 7)
2. **years_of_education** – počet let strávených vzděláváním
3. **faith** – víra (0 až 2)
4. **financial_situation** – finanční situace (1 až 4)
5. **income** – hrubá mzda (0 až 7)
6. **total_income** – souhrnný měsíční příjem (0 až 7)
7. **marital_status** – rodinný stav (0 až 3)
8. **student** – student (0 nebo 1)
9. **employee** – zaměstnanec (0 nebo 1)
10. **self_employed** – OSVČ (0 nebo 1)
11. **sex_m** – muž (0 nebo 1)
12. **sex_f** – žena (0 nebo 1)
13. **weight** – váha
14. **height** – výška
15. **age** – věk
16. **mental_health_hospitalized** – hospitalizace na psychiatrii (0 nebo 1)
17. **psychiatrist_help** – psychiatrická péče (0 až 2)
18. **psychologist_help** – psychologická péče (0 až 2)
19. **group_therapy** – skupinová terapie (0 nebo 1)
20. **psychology_training** – psychologické vzdělání/výcvik (0 až 3)
21. **mental_health_help_type** – psychologická/psychiatrická péče v posledním roce (0 až 3)
22. **diagnosed_adhd** – diagnostikováno ADHD (0 nebo 1)
23. **diagnosed_anxiety_depression** – diagnostikována úzkost/deprese (0 nebo 1)
24. **diagnosed_other** – diagnostikována jiná porucha (0 nebo 1)
25. **extroverted_enthusiastic** – extrovertní, nadšený/á (-3 až 3)
26. **critical_quarrelsome** – kritický/á, hádavý/á (-3 až 3)

27. **dependable_disciplined** – spolehlivý/á, disciplinovaný/á (-3 až 3)
28. **anxious_upset** – úzkostný/á, snadno rozrušitelný/á (-3 až 3)
29. **open_complex** – otevřený/á novým zkušenostem, všestranný (-3 až 3)
30. **reserved_quiet** – odměřený/á, tichý/á (-3 až 3)
31. **sympathetic_warm** – soucitný/á, srdečný/á (-3 až 3)
32. **disorganized_careless** – neorganizovaný/á, nedbalý/á (-3 až 3)
33. **calm_emotionally_stable** – klidný/á, emočně stabilní (-3 až 3)
34. **conventional_uncreative** – konvenční, nekreativní (-3 až 3)
35. **used_kratom** – užívání kratomu (0 nebo 1)
36. **used_alcohol** – užívání alkoholu (0 až 9)
37. **alcohol_addiction** – závislost na alkoholu (0 až 12)
38. **used_tobacco** – užívání tabáku (0 až 9)
39. **tobacco_addiction** – závislost na tabáku (0 až 12)
40. **used_cannabis** – užívání konopí (0 až 9)
41. **cannabis_addiction** – závislost na konopí (0 až 12)
42. **used_stimulants** – užívání stimulantů (0 až 9)
43. **stimulants_addiction** – závislost na stimulantech (0 až 12)
44. **used_opioids** – užívání opiátů (0 až 9)
45. **opioids_addiction** – závislost na opiátech (0 až 12)
46. **used_psychedelics** – užívání psychedelik (0 až 9)
47. **psychedelics_addiction** – závislost na psychedelikách (0 až 12)
48. **used_other_substances** – užívání jiných látek (0 až 9)
49. **other_substances_addiction** – závislost na jiných látkách (0 až 12)
50. **dependence_syndrome** – diagnostikován syndrom závislosti (0 nebo 1)
51. **addiction_hospitalized** – hospitalizace kvůli závislosti (0 nebo 1)
52. **addiction_outpatient** – ambulantní léčba se závislostí (0 nebo 1)
53. **past_year_psilocybin** – užívání psilocybinu v posledním roce (0 až 7)
54. **past_year_psilocybin_microdosing** – microdosing psilocybinu v posledním roce (0 nebo 1)
55. **past_year_lsd** – užívání LSD v posledním roce (0 až 7)
56. **past_year_lsd_microdosing** – microdosing LSD v posledním roce (0 nebo 1)

57. **past_year_mescaline** – užívání meskalinu v posledním roce (0 až 7)
58. **past_year_ayahuasca** – užívání ayahuasky v posledním roce (0 až 7)
59. **past_year_dmt** – užívání DMT v posledním roce (0 až 7)
60. **past_year_ketamine** – užívání ketaminu v posledním roce (0 až 7)
61. **past_year_mdma** – užívání MDMA v posledním roce (0 až 7)
62. **past_year_2cb** – užívání 2C-B v posledním roce (0 až 7)
63. **past_year_salvinorin** – užívání šalvěje divotvorné v posledním roce (0 až 7)
64. **past_year_muscimol** – užívání muchomůrky v posledním roce (0 až 7)
65. **past_year_others_microdosing** – microdosing jiných psychedelik v posledním roce (0 nebo 1)

Další skupina atributů hodnotí zejména setting během zkušenosti, motivace pro podstoupení zkušenosti a užité látky. Předpokládá se, že tyto údaje jsou známé ještě před samotným zážitkem:

1. **date_month** – měsíc (1 až 12)
2. **in_group** – užití ve skupině (0 nebo 1)
3. **p_home** – užití doma (0 nebo 1)
4. **p_nature** – užití v přírodě (0 nebo 1)
5. **p_music_event_bar** – užití v baru nebo na hudební akci (0 nebo 1)
6. **reason_to_use** – specifická motivace/důvod k užití (0 nebo 1)
7. **r_curiosity** – motivací je zvědavost (0 nebo 1)
8. **r_membership** – motivací je vyjádření sounáležitosti se sociální skupinou či přáteli (0 nebo 1)
9. **r_identity** – motivací je budování identity nebo pozornění na sebe (0 nebo 1)
10. **r_rebellion** – motivací je rebelie nebo alternativní životní styl (0 nebo 1)
11. **r_creativity** – motivací je stimulace uměleckého vyjádření a kreativity (0 nebo 1)
12. **r_pleasure** – motivací zintenzivněním prožitků a příjemných pocitů (0 nebo 1)
13. **r_social** – motivací zintenzivněním společenského kontaktu (0 nebo 1)
14. **r_compensation** – motivací je kompenzace vnitřního neuspokojení či nedostatečnosti (0 nebo 1)
15. **r_avoid_boredom** – motivací je vyhýbání se nudě či beznaději (0 nebo 1)

16. **r_enhance_mood** – motivací je zlepšení nálady (0 nebo 1)
17. **r_selfmedication** – motivací je samoléčba a zmírnění psychických či jiných problémů (0 nebo 1)
18. **r_selfknowledge** – motivací je sebepoznání (0 nebo 1)
19. **r_spiritual** – motivací jsou náboženské či spirituální důvody (0 nebo 1)
20. **music** – hudba (0 nebo 1)
21. **with_guide** – přítomnost průvodce (0 nebo 1)
22. **known_substance** – úroveň jistoty, o jakou jde látku (1 až 3)
23. **form_oral** – užití orálně (0 nebo 1)
24. **form_smoked** – užití kouřením (0 nebo 1)
25. **form_other** – užití jiným způsobem (0 nebo 1)
26. **subst_lsd** – užití LSD (0 nebo 1)
27. **subst_psilocybin** – užití psilocybinu (0 nebo 1)
28. **subst_cannabis** – užití konopí (0 nebo 1)
29. **subst_mdma** – užití MDMA (0 nebo 1)
30. **subst_alcohol** – užití alkoholu (0 nebo 1)
31. **subst_tobacco** – užití tabáku (0 nebo 1)
32. **subst_dmt** – užití DMT (0 nebo 1)
33. **subst_ketamine** – užití ketaminu (0 nebo 1)
34. **subst_ayahuasca** – užití ayahuasky (0 nebo 1)
35. **subst_2c_x** – užití 2C-B nebo příbuzné látky (0 nebo 1)
36. **subst_salvinorin_a** – užití šalvěje divotvorné (0 nebo 1)
37. **subst_mescaline** – užití meskalinu (0 nebo 1)
38. **subst_5meo_x** – užití 5-MeO-DMT nebo příbuzné látky (0 nebo 1)
39. **subst_stimulants** – užití stimulantů (0 nebo 1)
40. **subst_other** – užití jiné látky (0 nebo 1)
41. **subst_nonpharma** – nefarmakologické metody, například holotropní dýchání, tma a půsty (0 nebo 1)

Posledními příznaky v datovém souboru jsou již zmíněné třídy, jejichž hodnoty budou pomocí modelů predikovány:

1. **obn** – oceánská bezbřehost (*low*, *mid* nebo *high*)
2. **ded** – úzkost spojená s mizením ega (*low*, *mid* nebo *high*)
3. **vrs** – vizuální restrukturalizace (*low*, *mid* nebo *high*)
4. **edi** – míra rozpuštění ega (*low*, *mid* nebo *high*)

4.5 Vyhodnocování výkonnosti modelů

Aktuální dataset s 933 záznamy bude třeba rozdělit na dvě části – trénovací a testovací. První z nich bude použita v procesu učení modelů. Jednotlivé algoritmy v ní budou hledat vzorce a vztahy důležité pro predikci třídy. Poté bude výkonnost výsledných modelů experimentálně ověřována za pomoci testovacích dat. Volba rozdělení záleží na konkrétní situaci a množství dat, ale ve většině případů trénovací množina obsahuje 70 % až 90 % záznamů. Zkoumaný dataset není příliš velký a proto je navrženo rozdělení v podobě 733 (79,6 %) instancí pro trénink a 200 (21,4 %) pro testování. Autor zároveň předpokládá, že díky zvolenému poměru bude mít většina vypočítaných hodnot výkonnostních metrik ukončený desetinný rozvoj.

Výběr konkrétních způsobů, jak budou modely evaluovány, ovlivňuje několik předpokladů. Vzhledem k rozhodnutí vytvořit třídy pomocí rozřazení hodnot do tercilů je zřejmé, že instance budou rovnoměrně rozloženy mezi třídami. Zároveň platí, že není příliš velký rozdíl mezi jednotlivými druhy chyb.

Například, pokud při predikci DED bude skutečná hodnota v rámci *low*, ale model předpoví příslušnost k *high*, je to poměrně velký problém z důvodu, že záleží na počátečním setu člověka. Pokud si někdo myslí, že má velkou šanci mít úzkostný stav, reálně k němu může dojít kvůli zvýšené citlivosti během psychedelické zkušenosti, i když bez znalosti této chybné predikce by se pravděpodobně nedostavil. Je zřejmé, že v opačném případě, kdy je reálná hodnota DED v tercilu *high*, ale předpovídaný je *low*, jde o neméně závažnou chybu kvůli poskytnutí klamavé informace. To může mít za konečný důsledek zvýšení rizik. U dimenzí VRS a EDI navíc z definice nelze hodnotit příslušnost k jednotlivým třídám jako pozitivní nebo negativní.

Hlavní výkonnostní metrikou, jejíž hodnota bude zjišťována, je tedy **přesnost** (accuracy). Její použití je vhodné u vyvážených datasetů, jako je tento. Tato metrika ne-

rozlišuje mezi druhy chyb, ale v tomto kontextu jsou při vyhodnocování oba druhy považovány za srovnatelně špatné.

Klasifikační problém pracuje se třemi třídami a zároveň platí, že tyto kategorie jsou ordinální (lze určit jejich pořadí). Proto lze použít také **rozšířenou přesnost** (extended accuracy). Tato metrika poskytuje důležitou doplňkovou informaci k běžné přesnosti zejména u modelů, u kterých se předpokládá nižší přesnost z důvodu obecně špatné predikovatelnosti.

Běžné metriky jako preciznost, úplnost a F-míra nebudou použity. Přestože jde o časté způsoby evaluace výkonu v klasifikačních modelech, jsou vhodné spíše v situacích, kdy záleží na druhu chyby. Ve vícetřídní klasifikaci je navíc nutné preciznost a úplnost určovat pro každou třídu zvlášť, přičemž je z nich následně vypočítán aritmetický průměr, čímž dochází ke snížení výpovědní hodnoty získaného výsledku.

4.6 Optimalizace vstupních parametrů

Pro maximalizaci výkonu klasifikátorů je nutné vhodně definovat vstupní parametry. Ve většině případů platí, že čím komplexnější je algoritmus strojového učení, tím více vstupních parametrů akceptuje. Vzhledem k extrémnímu množství kombinací tohoto nastavení není možné je všechny otestovat, ale za pomoci různých aproximačních technik lze v relativně krátkém čase nalézt řešení, které se blíží optimu.

Tato práce k popsané optimalizační úloze přistupuje pomocí implementace genetického evolučního algoritmu využívajícího binárně definované chromozomy, což realizuje pomocí Pythonu. Existuje mnoho různých možností, jak GEA navrhnout, ale jak bylo popsáno v rámci teoretických východisek, vždy se jedná o kombinaci několika principů inspirovaných přírodou a biologickými strukturami.

Každý jedinec reprezentuje určité nastavení predikčního modelu. Hodnota **fitness funkce** tedy v této implementaci odpovídá přesnosti (accuracy) klasifikátoru, který je natrénován za použití parametrů představovaných tímto jedincem. Přestože tento konkrétní evoluční algoritmus pracuje výhradně s genomem ve dvojkové soustavě, ve skutečnosti je tato číselná řada přeložena před učením modelu do reálných čísel. Po-

čet bitů určených pro získání jednotlivých numerických hodnot v desítkové soustavě je předem známý.

Dalším důležitým principem v GEA je **křížení**. Zde je reprezentováno funkcí, která akceptuje dva rodičovské chromozomy A a B o stejné délce. Nejprve jsou vygenerovány masky M_1 a M_2 – dva náhodné řetězce složené z 0 a 1, které svou délkou odpovídají oběma vstupům. Maska determinuje, jaké bity potomek zdědí po kterém z rodičů. V genetických algoritmech platí, že pomocí jedné masky lze vytvořit dva nové chromozomy. V této implementaci je tedy za použití M_1 a M_2 možný vznik čtyřech jedinců.

K vytváření potomků jsou využívány binární operace AND a OR. Python nepodporuje binární NOT ve smyslu negace jednotlivých bitů v genomu, ale to lze nahradit pomocí binárního XOR (a za použití řetězce o délce vstupu složeného pouze z jedniček). Příklad 7 ukazuje vytvoření nových jedinců C, D, E a F z rodičů A a B za pomoci masek M_1 a M_2. Všechny použité logické operace jsou binární.

$$\begin{aligned} C &= (A \wedge M_1) \vee (B \wedge \neg M_1) \\ D &= (B \wedge M_1) \vee (A \wedge \neg M_1) \\ E &= (A \wedge M_2) \vee (B \wedge \neg M_2) \\ F &= (B \wedge M_2) \vee (A \wedge \neg M_2) \end{aligned} \tag{7}$$

V každém z nově vytvořených jedinců může dojít k následné **mutaci** s pravděpodobností 25 %. V takovém případě je zaměněn jeden náhodný bit v řetězci za opačnou hodnotu. Mutace je důležitá k zejména z hlediska eliminace rizika stagnace v lokálním maximu.

Ve všech generacích jsou jedinci vybíráni k dalšímu křížení pomocí **selekce podle pořadí**. Jsou tedy v rámci pole seřazeni v závislosti na hodnotě fitness funkce, přičemž pravděpodobnost výběru jedince je úměrná této pozici. Nejvyšší pravděpodobnost výběru má tedy poslední v poli. Poté je pro další křížení selektováno náhodné sudé množství jedinců bez opakování, avšak minimálně čtvrtina a maximálně tři čtvrtiny populace. Z nich jsou utvořeny páry (první s druhým, třetí se čtvrtým, atd.) a z

každé takové dvojice vzniknou čtyři potomci, kteří vstoupí do nové generace. Selekce podle pořadí je vhodná zejména z důvodu navržené fitness funkce, která v praxi vrací poměrně podobné hodnoty pro jednotlivé jedince, což by těm nejlepším přineslo příliš malé zvýhodnění.

Před spuštěním genetického evolučního algoritmu je zapotřebí definovat hyperparametry, které ovlivňují jeho následný běh a optimalizační proces. Jsou to konkrétně:

- fitness funkce, která se liší podle zkoumaného algoritmu a pro jedince (nastavení) vrací přesnost natrénovaného modelu,
- pole přirozených čísel reprezentující počty bitů (rozsah hodnot) pro každý z optimalizovaných vstupních parametrů klasifikátoru,
- velikost počáteční populace,
- maximální počet generací.

Na základě hyperparametrů je nejprve vygenerována počáteční populace. Pro všechny jedince je vypočítána hodnota fitness funkce, podle níž jsou seřazeni v poli. Umístění je zásadní pro vstup do křížení, do kterého jsou vybráni nejpravděpodobněji nejlepší jedinci. Proces křížení každých dvou genomů vytvoří čtyři nové jedince, kteří budou součástí nadcházející generace. Tento proces se opakuje do doby, kdy je buď dosažen maximální počet generací, nebo jedinců zbývá příliš málo. Algoritmus průběžně zaznamenává nejlepší genomy v každé generaci, kteří představují nejkvalitnější nalezená nastavení parametrů pro vybraný model strojového učení.

4.7 Výběr atributů

Zvýšení efektivity učení a celkové výkonnosti modelů lze dosáhnout pomocí vhodného výběru atributů. Jak vyplývá z teoretických východisek práce, existuje mnoho způsobů provedení této selekce, z nichž některé jsou již součástí implementace algoritmů strojového učení. Přesto je vhodné před jejich použitím eliminovat atributy, které pravděpodobně nemají vliv na přesnost predikcí. Kromě teoreticky vyšší výkonnosti je počáteční odstranění některých příznaků výhodné z hlediska případného nasazení modelu v praxi, neboť se snižuje počet vstupů, které uživatel musí zadat, aby získal predikci pro svůj případ.

Jedním z nejběžnějších způsobů, jak lze zjistit provázanost dvou proměnných je korelace. Při použití Pearsonova korelačního koeficientu lze získat hodnotu v intervalu $\langle -1, 1 \rangle$, která ukazuje sílu závislosti od zcela nepřímé (-1) po zcela přímou (1). Pokud žádná lineární závislost neexistuje, je koeficient roven nule.

Aby bylo dosaženo přesnějších výsledků korelačního koeficientu, není příliš vhodné pro výpočet používat predikované (závislé) proměnné rozdělené do tercilů. Kvalitnější údaje je možné získat, pokud bude síla korelace zjišťována oproti původním proměnným s hodnotami v intervalu $\langle 0, 100 \rangle$.

Pro každou kombinaci nezávislé a predikované proměnné lze tedy vypočítat sílu této korelace, tedy vliv konkrétního příznaku na výslednou předpověď. Evoluční algoritmus umožňuje mezi optimalizované parametry zařadit také minimální úroveň korelace (respektive její absolutní hodnotu). Pro natrénování modelu (a potažmo určení hodnoty fitness funkce) bude tedy nutné vždy předem upravit dataset a eliminovat příznaky, které mají slabší korelaci s predikovanou proměnnou, než je předem určeno.

Tento přístup spíše nepatří mezi běžné, ale ve zkoumaném datasetu jsou téměř všechny zjištěné korelace velmi slabé. Není tedy příliš snadné předem racionálně určit hranici pro minimální hodnotu Pearsonova korelačního koeficientu pro ponechání ve vstupním souboru. Příznaky s nejsilnější korelací vůči jednotlivým predikovaným proměnným zobrazuje tabulka 7, z níž je zřejmé, že ze 106 proměnných:

- s OBN slabě koreluje 5 proměnných a 101 velmi slabě,
- s DED slabě koreluje 1 proměnná a 105 velmi slabě,
- s VRS slabě koreluje 7 proměnných a 99 velmi slabě,
- s EDI slabě koreluje 1 proměnná a 105 velmi slabě.

Platí, že pro většinu (zejména složitějších) predikčních modelů je typické, že nejsou pro člověka příliš interpretovatelné – často tedy nelze analyzovat rozhodnutí, která vedla ke konkrétní předpovědi. Tato práce uvádí přehled příznaků s nejsilnější korelací (tabulka 7) i z důvodu, že jde o vhodný komplement k těmto klasifikátorům. Ukazuje tak pravděpodobný vliv proměnných na predikci, čímž pomáhá zvyšovat úroveň informovanosti potenciálních uživatelů.

Oceánská bezbřehost	
r_spiritual	0.228
sex_m	-0.218
sex_f	0.213
conventional_uncreative	-0.208
r_selfknowledge	0.207
used_stimulants	0.189
past_year_mdma	0.188
used_psychedelics	0.179
r_identity	0.164
past_year_psilocybin	0.160
used_tobacco	0.156
past_year_lsd	0.151

Úzkost spojená s mizením ega	
diagnosed_adhd	0.241
r_rebellion	0.189
alcohol_addiction	0.184
dependable_disciplined	-0.182
r_identity	0.164
calm_emotionally_stable	-0.147
r_avoid_boredom	0.139
used_tobacco	0.138
mental_health_hospitalized	0.137
subst_salvinorin_a	0.135
used_opioids	0.131
past_year_salvinorin	0.128

Vizuální restrukturalizace	
r_identity	0.239
subst_lsd	0.234
r_selfknowledge	0.222
used_tobacco	0.217
used_psychedelics	0.208
alcohol_addiction	0.205
past_year_psilocybin	0.199
used_stimulants	0.184
past_year_mdma	0.180
past_year_lsd	0.178
form_oral	0.175
form_smoked	-0.171

Míra rozpuštění ega	
r_spiritual	0.195
r_selfknowledge	0.168
used_tobacco	0.164
sex_m	-0.163
past_year_psilocybin	0.161
used_stimulants	0.160
conventional_uncreative	-0.159
sex_f	0.148
used_psychedelics	0.140
past_year_lsd	0.138
past_year_mdma	0.138
with_guide	0.133

Tabulka 7: Příznaky s nejsilnější korelací vůči predikovaným proměnným

Zdroj: vlastní zpracování

4.8 Predikční modely

Tato část práce se zabývá zkoumáním několika vybraných predikčních modelů, které řeší klasifikaci psychedelické zkušenosti do jednoho ze tří tercilů v rámci každé ze čtyř dříve představených dimenzí. Praktická implementace bude realizována pomocí jazyka Python a knihoven jako Scikit-learn, NumPy a Keras.

Pro optimalizaci počátečního nastavení bude využit připravený genetický evoluční algoritmus. Snížení počtu takto aproximovaných parametrů (zmenšení prostoru možných řešení) zvyšuje pravděpodobnost nalezení lepších výsledků. Z toho důvodu je vhodné konfiguraci částečně připravit předem. Některé z parametrů budou tedy v průběhu optimalizačního procesu považovány za konstantní a budou definovány pomocí často využívaných, doporučených či výchozích hodnot. GEA pracuje s počáteční populací 20 náhodných jedinců a počet generací se liší podle výpočetní složitosti jednotlivých algoritmů.

Jak již bylo popsáno, každý z klasifikátorů pracuje s datasetem, jehož příznaky respektují minimální absolutní hodnotu síly korelace s predikovanou proměnnou. Aby bylo jednodušší v rámci GEA optimalizovat tento parametr a odpadla nutnost pracovat s desetinnými hodnotami, je využito přirozené číslo n. Z něj se konkrétní minimum Pearsonova korelačního koeficientu pro jednotlivé proměnné spočítá jako $\frac{n}{1000}$.

4.8.1 Rozhodovací strom

Rozhodovací strom je sestaven pomocí DecisionTreeClassifier z knihovny Scikit-learn. Běh tohoto druhu algoritmu je determinován především maximální hloubkou stromu (max_depth). Důležitým parametrem, který snižuje riziko overfittingu je minimální počet vzorků na listech (min_samples_leaf). Rozdělení tedy proběhne pouze v případě, že v každé z nových větví bude zachováno alespoň toto minimální množství instancí. Implementace ve Scikit-learn umožňuje také definovat seed (random_state), který je využit v rámci generátoru náhodných čísel a zároveň zajistí determinističnost napříč jednotlivými běhy.

Tyto tři parametry, doplněné o minimální sílu korelace příznaku pro zachování v datasetu, jsou optimalizovány pomocí evolučního algoritmu. Nejlepší nalezená konfigurace

pro jednotlivé predikční modely je prezentována v tabulce 8. Rozsah značí počet možných hodnot vycházející z vyhrazených bitů v genomu. Rozhodovací stromy jsou poměrně efektivní a je tedy možné vytvořit velké množství jedinců a populací. Výkonnost se ale přestala zlepšovat už s 16. generací.

Parametr	Rozsah	OBN	DED	VRS	EDI
Min. korelace (abs. hodnota)	256	0.062	0.025	0.059	0.046
random_state	512	452	54	190	67
min_samples_leaf	32	1	1	1	1
max_depth	128	96	100	84	107

Tabulka 8: Nejlepší nalezená nastavení pro rozhodovací strom

Zdroj: vlastní zpracování

Hodnoty přesnosti a rozšířené přesnosti pro tato nejlepší nalezená nastavení jsou uvedena v tabulce 9 pro každou predikovanou proměnnou.

Proměnná	Přesnost	Rozšířená přesnost
OBN	65.5 %	95 %
DED	63 %	92.5 %
VRS	66 %	95 %
EDI	64.5 %	92 %

Tabulka 9: Výkonnostní metriky pro rozhodovací strom

Zdroj: vlastní zpracování

4.8.2 Náhodný les

Náhodný les je implementován pomocí RandomForestClassifier ze stejné knihovny. Hlavním nastavením, které je třeba optimalizovat, je počet stromů použitých v rámci modelu (n_estimators). V tomto případě je také možné definovat seed (random_state) pro využití v generátoru náhodných čísel. Přestože náhodný les nabízí mnoho možností konfigurace (zejména pro jednotlivé stromy), během testování se ukázalo, že jiné než výchozí parametry mají spíše negativní vliv na výkonnost výsledných modelů.

Dva vybrané parametry pro RandomForestClassifier jsou opět optimalizovány spolu s minimální sílou korelace a vypsány v tabulce 10. Uveden je i počet možných hodnot (rozsah). Náhodný les není příliš rychlý v případech, kdy je použit vysoký počet estimátorů (stromů). Běh byl proto omezen 15 generacemi, přičemž ve dvou posledních měli nejlepší jedinci poměrně podobnou hodnotu fitness funkce.

Parametr	Rozsah	OBN	DED	VRS	EDI
Min. korelace (abs. hodnota)	256	0.004	0.014	0.007	0.046
random_state	512	128	46	167	89
n_estimators	1024	233	688	328	515

Tabulka 10: Nejlepší nalezená nastavení pro náhodný les

Zdroj: vlastní zpracování

Konkrétní výsledky z hlediska výkonnostních metrik jsou prezentovány v tabulce 11 pro všechny čtyři predikované proměnné.

Proměnná	Přesnost	Rozšířená přesnost
OBN	71.5 %	95.5 %
DED	69.5 %	93 %
VRS	73.5 %	96 %
EDI	69 %	91.5 %

Tabulka 11: Výkonnostní metriky pro náhodný les

Zdroj: vlastní zpracování

4.8.3 Naive Bayes

Tato práce využívá gaussovský Naive Bayes (GaussianNB) taktéž ze zmíněné knihovny Scikit-learn. V této implementaci nejsou žádné parametry, které by bylo výhodné optimalizovat. Vzhledem k tomu, že Naive Bayes je zároveň velmi rychlý algoritmus, je možné vyzkoušet všechny absolutní hodnoty minimální síly korelace, které jsou zastoupeny v datasetu. Nejlepší nastavení pro jednotlivé proměnné jsou ukázána v tabulce 12.

Parametr	OBN	DED	VRS	EDI
Min. korelace (abs. hodnota)	0.1	0.048	0.081	0.049

Tabulka 12: Nejlepší nalezená nastavení pro Naive Bayes

Zdroj: vlastní zpracování

Výsledky výkonnostních metrik pro gaussovský Naive Bayes a jednotlivé proměnné jsou prezentovány v tabulce 13.

Proměnná	Přesnost	Rozšířená přesnost
OBN	64.5 %	94 %
DED	63 %	91 %
VRS	64.5 %	93.5 %
EDI	61.5 %	92.5 %

Tabulka 13: Výkonnostní metriky pro Naive Bayes

Zdroj: vlastní zpracování

4.8.4 Algoritmus podpůrných vektorů

Algoritmus podpůrných vektorů využívá implementaci SVC ze Scikit-learn, která pracuje s regularizačním parametrem C. Během testování se ale ukázalo, že není možné model za reálnou dobu natrénovat pro jiný typ jádra, než lineární. Z toho důvodu nebude optimalizováno nastavení související s jinými typy jader, jako je například stupeň polynomu. Zároveň je nutné zohlednit výpočetní náročnost při vysokých hodnotách regularizačního parametru C a omezit prostor možných řešení. Vzhledem k získání žádoucího rozsahu je použita vždy desetina z hodnoty C (princip výpočtu je stejný, jako byl popsán u výběru atributů). Evoluční algoritmus i v tomto případě řešil minimální úroveň korelace a z důvodu pomalejších výpočtů běžel pouze po 10 generací. Nejlepší konfigurace je prezentována v tabulce 14.

Hodnoty přesnosti a rozšířené přesnosti pro tato nejlepší nalezená nastavení jsou uvedena v tabulce 15 pro každou predikovanou proměnnou.

Parametr	Rozsah	OBN	DED	VRS	EDI
Min. korelace (abs. hodnota)	256	0.012	0.009	0.017	0.006
C	64	4.0	5.1	3.2	4.1

Tabulka 14: Nejlepší nalezená nastavení pro algoritmus podpůrných vektorů

Zdroj: vlastní zpracování

Proměnná	Přesnost	Rozšířená přesnost
OBN	61.5 %	89.5 %
DED	59 %	90.5 %
VRS	62 %	91.5 %
EDI	59.5 %	92 %

Tabulka 15: Výkonnostní metriky pro algoritmus podpůrných vektorů

Zdroj: vlastní zpracování

4.8.5 Neuronová síť

Neuronové sítě obecně patří mezi nejvýkonnější predikční modely. Jedná se o velmi komplexní struktury se širokými možnostmi konfigurace pro různé problémy. Zároveň platí, že trénovací proces často trvá poměrně dlouho. Z těchto důvodů je nutné síť zčásti navrhnout před začátkem optimalizačního procesu a pracovat s co nejmenším prostorem řešení.

Pro tuto klasifikační úlohu je navržena plně propojená sekvenční síť s jednou skrytou vrstvou pomocí knihovny Keras. Ve vstupní a skryté vrstvě je použita aktivační funkce ReLU. Na výstupu jsou tři neurony využívající softmax, díky čemuž je možné získat pravděpodobnost příslušnosti ke každému ze tří tercilů. Velikost chyby zjišťuje široce využívaná kategorická křížová entropie pracující s metrikou přesnosti.

Tato architektura byla zvolena na základě obecných doporučení pro návrh neuronových sítí popsaných v teoretické části práce. Je využívána pouze jedna skrytá vrstva z důvodu spíše nižší komplexnosti problému. Vzhledem ke způsobu implementace vyhodnocování fitness funkce v genetickém algoritmu je také vhodné dosáhnout co nejvyšší rychlosti učení, na což má počet vrstev zásadní vliv.

V evolučním algoritmu jsou tedy optimalizovány tyto parametry:

1. absolutní hodnota minimální síly korelace pro ponechání příznaku v souboru,
2. počet neuronů ve vstupní vrstvě,
3. počet neuronů ve skryté vrstvě,
4. rychlost učení (learning_rate),
5. počet epoch – kolikrát síť projde celý vstupní soubor (epochs),
6. velikost dávky – počet vzorků šířených skrz síť (batch_size).

Vzhledem k náročnosti výpočtů bylo pro GEA zvoleno 10 generací. Nejlepší nalezené konfigurace pro OBN, DED, VRS a EDI jsou uvedeny v tabulce 16.

Parametr	Rozsah	OBN	DED	VRS	EDI
Min. korelace (abs. hodnota)	128	0.058	0.029	0.04	0.047
Počet neuronů – vstupní v.	128	102	101	123	31
Počet neuronů – skrytá v.	256	169	176	42	96
learning_rate	32	0.01	0.014	0.027	0.017
epochs	256	117	230	168	179
batch_size	256	40	129	88	137

Tabulka 16: Nejlepší nalezená nastavení pro neuronovou síť

Zdroj: vlastní zpracování

Výsledky z hlediska výkonnostních metrik pro natrénovanou neuronovou síť jsou prezentovány v tabulce 17 pro všechny čtyři predikované proměnné.

Proměnná	Přesnost	Rozšířená přesnost
OBN	74 %	96 %
DED	71.5 %	94.5 %
VRS	73 %	94 %
EDI	71 %	95 %

Tabulka 17: Výkonnostní metriky pro neuronovou síť

Zdroj: vlastní zpracování

5 Výsledky a diskuze

5.1 Výkonnost a vlastnosti modelů

Výkonnost klasifikačních modelů lze porovnávat z vícero pohledů. Základní metrikou použitou k určení kvality modelů v této práci byla přesnost. V tomto ohledu byla pro oceánskou bezbřehost (OBN), úzkost spojenou s mizením ega (DED) a míru rozpuštění ega (EDI) nejlepším modelem neuronová síť. Vizuální restrukturalizaci (VRS) nejpřesněji předvídal náhodný les. Tyto výsledky jsou prezentovány v tabulce 18.

Algoritmus	OBN	DED	VRS	EDI
Rozhodovací strom	65.5 %	63 %	66 %	64.5 %
Náhodný les	71.5 %	69.5 %	73.5 %	69 %
Naive Bayes	64.5 %	63 %	64.5 %	61.5 %
Algoritmus podpůrných vektorů	61.5 %	59 %	62 %	59.5 %
Neuronová síť	74 %	71.5 %	73 %	71 %

Tabulka 18: Porovnání přesnosti modelů pro všechny predikované proměnné

Zdroj: vlastní zpracování

Lze říci, že implementovaná neuronová síť je z hlediska hodnot přesnosti i rozšířené přesnosti obecně nejlepším klasifikátorem pro tento problém. Mezi její hlavní nevýhody ale patří výpočetní složitost pro trénování, která v důsledku snížila počet kombinací počátečního nastavení, které je možné v určitém čase otestovat. To tedy může znamenat, že při vhodnější konfiguraci by poskytovala ještě lepší výstupy.

Predikční model využívající náhodný les by bylo možné označit za druhý nejlepší. Jeho výkonnost je podobná neuronové síti s ohledem na přesnost i rozšířenou přesnost. Oproti ní má však velkou výhodu v rychlosti učení a v nižším počtu parametrů, což umožnilo vyzkoušet násobně více možností konfigurace. Je pravděpodobné, že jeho výkonnost je velmi blízko optimu.

Náhodný les a neuronová síť mají společnou vlastnost – kromě samotné predikce je známá i její pravděpodobnost. Implementovaná síť má u korektně určených predikcí průměrnou pravděpodobnost 96.6 %, zatímco u špatných 88.03 %. U náhodného lesa

je tento poměr odlišný – správné předpovědi mají 71.2 % a chybné 49.1 %. V tomto ohledu je tedy druhý zmíněný klasifikátor výrazně úspěšnější, protože poskytuje mnohem reálnější odhady.

Výkonnost Naive Bayes z hlediska přesnosti není tak vysoká, jako u náhodného lesa a neuronové sítě, ale je s nimi srovnatelný z pohledu rozšířené přesnosti. Pro svoje fungování ale vyžaduje nejméně atributů ze všech testovaných klasifikátorů, což při nasazení do praxe šetří čas lidem při vyplňování vstupních parametrů – např. pro klasifikaci OBN jich vyžaduje 48, což je necelá polovina původního datasetu.

Jednoduchý rozhodovací strom patřil podle původních očekávání kvůli své nízké komplexnosti mezi horší modely. Díky extrémně rychlému učícímu procesu je velká pravděpodobnost, že nalezené řešení je optimální. Během aproximace pomocí evolučního algoritmu již nevykazoval žádná zlepšení v posledních generacích.

Algoritmus podpůrných vektorů (SVC) byl jednoznačně nejhorší. Kromě velmi slabého výkonu jeho učení trvalo dlouhou dobu a bylo neefektivní v porovnání s ostatními algoritmy. V kontextu tohoto klasifikačního problému se tedy jedná o nevhodné řešení.

5.1.1 Optimalizace počáteční konfigurace

Pro hledání vhodného počátečního nastavení byl navržen a implementován evoluční genetický algoritmus, který lze hodnotit jako efektivní. Zvláště pro jednodušší klasifikační modely byla aproximace vstupních parametrů velmi rychlá a v prvních generacích se výrazně zlepšovala hodnota fitness funkce nejlepších jedinců.

Přestože způsob volby vhodné podmnožiny příznaků na základě korelace má určitá omezení, v rámci evolučního algoritmu se tento přístup prokázal jako poměrně dobře zvolený zejména z důvodu rychlosti. To je klíčový parametr vzhledem k nutnosti testování vysokého počtu různých nastavení.

Využití genetického algoritmu pro počáteční konfiguraci neuronové sítě nicméně nelze označit za vhodné. Pro získání hodnoty fitness funkce je nutné natrénovat a otestovat celý model, což je v tomto případě poměrně výpočetně náročné. Pravděpodobným lep-

ším řešením by bylo nalezení efektivnějšího způsobu definování fitness funkce v evolučním algoritmu. Alternativou je arbitrární určení parametrů sítě a trénování s využitím validačního datasetu.

5.2 Hodnocení naplnění cílů práce

Lze konstatovat, že hlavní cíl, tedy nalezení vhodného způsobu, jak je možné využít strojové učení pro předpovídaní charakteru psychedelické zkušenosti, byl naplněn za využití několika klasifikačních modelů. Následné experimentální ověření nicméně prokázalo, že jsou tyto zážitky poněkud špatně predikovatelné.

Počáteční předpoklad spíše nižší přesnosti modelů byl tedy potvrzen, přestože proběhlo množství úprav dat ve vstupním souboru, které obecně vedou ke zlepšování výkonů při využití strojového učení. Je tedy velmi pravděpodobné, že psychedelickou zkušenost značně ovlivňují i jiné aspekty setu a settingu, jejichž exaktní popis a kvantifikaci není snadné pomocí dotazníkového šetření získat (například přesné údaje o psychickém stavu). Přesnost predikce mohla být také negativně ovlivněna přítomností šumu a nepravdivých údajů v datasetu.

V současné době existuje jen velmi málo akademických prací, které se zabývají předpovídáním jednotlivých aspektů psychedelické zkušenosti pomocí strojového učení. Není tedy zatím možné dosažené výsledky porovnat s jinými publikacemi. Výzkum psychedelik je ale během posledních let na vzestupu, takže je pravděpodobné, že tato práce získá v budoucnu nový kontext, v rámci kterého bude možné lépe vyhodnotit kvalitu natrénovaných modelů.

5.3 Využití v praxi

I přes popsané nedostatky lze klasifikátory ale považovat dostatečně přesné, aby mohly nalézt praktické využití pro snížení rizik spojených s psychedelickou zkušeností. V případě takového nasazení by bylo žádoucí uvést i informaci o dosažených hodnotách výkonnostních metrik, aby potenciální uživatelé nebyli klamavě informováni. Vzhledem k tomu, že z komplexnějších modelů není kvůli nižší interpretovatelnosti zřejmé, jakým způsobem bylo konkrétní předpovědi dosaženo, bylo by vhodné dodat i přehled síly

korelace jednotlivých proměnných s OBN, DED, VRS a EDI.

Pravděpodobně nejlepší volbou pro použití v praxi by byl podle názoru autora predikční model pracující s algoritmem náhodného lesa. Přestože je o něco méně přesný v porovnání s neuronovou sítí, nabízí lepší informace z hlediska pravděpodobnosti správnosti získané předpovědi, což by mohl být pro potenciální uživatele poměrně cenný údaj.

Při využití v praxi je ale nutné důsledně dbát na obecná pravidla bezpečného užívání psychedelických látek a zejména na individuální aspekty, které není možné plně zohlednit v jakémkoli predikčním modelu.

5.4 Možnosti zlepšení a další práce

Je velmi pravděpodobné, že výkonnost modelu by mohla být zlepšena při navýšení počtu příznaků popisujících psychický stav, individuální představy či očekávání, která mají lidé před zahájením zážitku. Pozitivní vliv by nejspíše mělo i získání většího množství instancí.

Další práce s využitím datasetu v aktuální podobě by mohla zahrnovat změnu implementace genetického evolučního algoritmu pro použití v neuronových sítích. To by znamenalo zejména obměnu fitness funkce, protože v současné podobě je získání její hodnoty poměrně výpočetně náročné. Evoluční algoritmus by případně vůbec nemusel být použit. Očekávaným výsledkem by bylo zvýšení výkonnosti tohoto typu predikčního modelu.

Vzhledem k uspokojivým výsledkům z hlediska pravděpodobnosti správnosti předpovědi u náhodného lesa by také bylo možné nepříliš jisté předpovědi přehodnotit například pomocí neuronové sítě. Využití tohoto přístupu je ale podmíněno tím, že predikční algoritmy chybují ve většině případů u jiných instancí, což by bylo nutné před implementací ověřit.

6 Závěr

Tato práce se zabývala hledáním způsobu, jak by bylo možné využít strojové učení pro snížení rizik při užívání psychedelických látek. Vychází z předpokladu, že případné negativní dopady činností je možné minimalizovat pomocí dobré informovanosti.

Nejprve byla formulována teoretická východiska pro psychedelické zkušenosti, jednotlivé látky a jejich využití. Zároveň byl ukázán způsob, jak lze tyto zážitky kvantifikovat z různých hledisek. Důležitým zjištěním v této oblasti bylo, že riziko fyzických dopadů užívání a vzniku závislosti je zanedbatelné, ale je zásadní dbát na správnou souhru myšlenkového nastavení (set) a prostředí, kde se tato zkušenost odehrává (setting). Tyto parametry mají podstatný vliv na průběh zážitku a v případě jejich nevhodné kombinace se může objevit náročná zkušenost, která je typická nepříjemnými psychickými pocity a úzkostmi.

V rámci teoretické části práce byly také studovány různé algoritmy strojového učení, spolu s jejich kladnými a zápornými stránkami. Jsou uvedeny způsoby, jak lze sestavit predikční (klasifikační) modely a následně měřit jejich výkonnost. Byly popsány způsoby jejich počáteční konfigurace a optimalizace parametrů. Práce zde také uvádí vhodné přístupy k úpravám vstupního datasetu vzhledem k maximalizaci efektivity učení a výkonnosti modelů.

Praktická část v úvodu popisuje zdroj dat a metody jejich zpracování pro následné použití k trénování klasifikátorů. Dataset byl upraven zejména tak, aby v něm nechyběly žádné hodnoty a všechny údaje byly numerické. Na základě zkoumaného problému byla formulována predikční klasifikační úloha a identifikovány předpovídané proměnné (oceánská bezbřehost, úzkost spojená s mizením ega, vizuální restrukturalizace a míra rozpuštění ega). Bylo využito rozdělení původně numerických predikovaných proměnných na tercily a popsány důvody výhodnosti tohoto rozhodnutí. Práce uvádí kompletní přehled atributů v souboru po provedených úpravách a jejich popis (datový slovník).

Poté byly představeny způsoby měření výkonnosti výsledných modelů (přesnost a rozšířená přesnost). Dále byl uveden způsob optimalizace počátečního nastavení klasifi-

kátorů – genetický evoluční algoritmus, jehož implementace byla detailně vysvětlena. Výběr atributů byl proveden pomocí definování minimální absolutní hodnoty síly korelace jednotlivých příznaků v souboru s predikovanými proměnnými (pomocí Pearsonova korelačního koeficientu). Hledání této minimální úrovně bylo taktéž realizováno za použití aproximace evolučním algoritmem.

Každý z pěti natrénovaných modelů strojového učení byl otestován a ohodnocen pomocí předem zvolených výkonnostních metrik. Zároveň byla uvedena nejlepší možná konfigurace nalezená pomocí genetického algoritmu, při které modely dosahovaly prezentovanou výkonnost.

V závěru práce prezentuje, analyzuje a shrnuje získané výsledky z různých pohledů a jako nejlepší modely uvádí ty, které využívají neuronovou síť a náhodný les. Byly ukázány důvody, proč je evoluční algoritmus vhodný pro optimalizaci konfigurace u jednodušších algoritmů a naopak přináší problémy při využití v neuronové síti. Bylo zhodnoceno naplnění cílů práce, možnosti využití modelů v praxi. Taktéž bylo ukázáno, kde je prostor pro zlepšení a další rozvoj.

7 Seznam použitých zdrojů

[1] OSMOND, Humphry. A Review of the Clinical Effects of Psychotomimetic Agents. Annals of the New York Academy of Sciences. 1957, 66(3), 418-434. ISSN 00778923. Dostupné z: doi:10.1111/j.1749-6632.1957.tb40738.x

[2] JOHNSON, Matthew W., Peter S. HENDRICKS, Frederick S. BARRETT a Roland R. GRIFFITHS. Classic psychedelics: An integrative review of epidemiology, therapeutics, mystical experience, and brain network function. Pharmacology & Therapeutics. 2019, 197, 83-102 . ISSN 01637258. Dostupné z: doi:10.1016/j.pharmthera.2018.11.010

[3] GROF, Stanislav. Kosmická hra: zkoumání hranic lidského vědomí. Praha: Perla, 1998. ISBN 80-902156-1-0.

[4] BIGELOW, Barbara C. a Kathleen J. EDGAR. UXL Encyclopedia of Drugs and Addictive Substances. UXL, 2006. ISBN 9781414404486.

[5] O'BRIEN, Charles P. Drug addiction and drug abuse. Goodman and Gilman's the pharmacological basis of therapeutics, 2006, 11: 607-627. Dostupné z: http://ndl.ethernet.edu.et/bitstream/123456789/19620/1/1780.pdf

[6] HOFMANN, Albert. LSD, my problem child. New York: McGraw-Hill, 1980. ISBN 0-07-029325-2.

[7] TYLŠ, Filip, Tomáš PÁLENÍČEK a Jiří HORÁČEK. Psilocybin – Summary of knowledge and new perspectives. European Neuropsychopharmacology. 2014, 24(3), 342-356. ISSN 0924977X. Dostupné z: doi:10.1016/j.euroneuro.2013.12.006

[8] POSTRÁNECKÁ, Zuzana, Čestmír VEJMOLA a Filip TYLŠ. Psychedelic therapy in the Czech Republic: A theoretical concept or a realistic goal?. Journal of Psychedelic Studies. 2019, 3(1), 19-31. ISSN 2559-9283. Dostupné z: doi:10.1556/2054.2019.003

[9] TYLŠ, Filip. Fenomén psychedelie: subjektivní popisy zážitků z experimentální intoxikace psilocybinem doplněné pohledy výzkumníků. Vydání druhé. Praha: Dy-

bbuk, 2020. ISBN 978-80-7438-226-0.

[10] GABLE, Robert S. Risk assessment of ritual use of oral dimethyltryptamine (DMT) and harmala alkaloids. Addiction. 2007, 102(1), 24-34. ISSN 09652140. Dostupné z: doi:10.1111/j.1360-0443.2006.01652.x

[11] ERMAKOVA, Anna O, Fiona DUNBAR, James RUCKER a Matthew W JOHNSON. A narrative synthesis of research with 5-MeO-DMT. Journal of Psychopharmacology. 2022, 36(3), 273-294. ISSN 0269-8811. Dostupné z: doi:10.1177/02698811211050543

[12] VAMVAKOPOULOU, Ioanna A., Kelly A.D. NARINE, Ian CAMPBELL, Jason R.B. DYCK a David J. NUTT. Mescaline: The forgotten psychedelic. Neuropharmacology [online]. 2023, 222 [cit. 2023-03-11]. ISSN 00283908. Dostupné z: doi:10.1016/j.neuropharm.2022.109294

[13] KOENIG, Xaver a Karlheinz HILBER. The Anti-Addiction Drug Ibogaine and the Heart: A Delicate Relation. Molecules. 2015, 20(2), 2208-2228. ISSN 1420-3049. Dostupné z: doi:10.3390/molecules20022208

[14] ANDRASHKO, Veronika, Tomas NOVAK, Martin BRUNOVSKY, Monika KLIROVA, Peter SOS a Jiri HORACEK. The Antidepressant Effect of Ketamine Is Dampened by Concomitant Benzodiazepine Medication. Frontiers in Psychiatry. 2020, 11. ISSN 1664-0640. Dostupné z: doi:10.3389/fpsyt.2020.00844

[15] CÉSAROVÁ, Eva. Možnosti a meze alternativní léčby a úzdravy ze závislosti prostřednictvím psychedelické zkušenosti. Praha, 2021. Diplomová práce. Univerzita Karlova v Praze, 1. lékařská fakulta.

[16] NUGTEREN-VAN LONKHUYZEN, Johanna J., Antoinette J.H.P. VAN RIEL, Tibor M. BRUNT a Laura HONDEBRINK. Pharmacokinetics, pharmacodynamics and toxicology of new psychoactive substances (NPS): 2C-B, 4-fluoroamphetamine and benzofurans. Drug and Alcohol Dependence. 2015, 157, 18-27. ISSN 03768716. Dostupné z: doi:10.1016/j.drugalcdep.2015.10.011

[17] HATIPOGLU, Seda Damla, Burhanettin YALCINKAYA, Muslum AKGOZ, Turan OZTURK, Ahmet C. GOREN a Gulacti TOPCU. Screening of Hallucinogenic Compounds and Genomic Characterisation of 40 Anatolian Salvia Species. Phytochemical Analysis. 2017, 28(6), 541-549. ISSN 09580344. Dostupné z: doi:10.1002/pca.2703

[18] HARDING, Wayne W., Matthew SCHMIDT, Kevin TIDGEWELL, et al. Synthetic Studies of Neoclerodane Diterpenes from Salvia divinorum: Semisynthesis of Salvinicins A and B and Other Chemical Transformations of Salvinorin A. Journal of Natural Products. 2006, 69(1), 107-112. ISSN 0163-3864. Dostupné z: doi:10.1021/np050398i

[19] RÄTSCH, Christian. Psychoaktivní rostliny: historie, léčení, účinky, příprava, rituály. Olomouc: Fontána, [2011]. ISBN 978-80-7336-625-4.

[20] VAN WEL, JHP, DB SPRONK, KPC KUYPERS, EL THEUNISSEN, SW TOENNES, RJ VERKES a JG RAMAEKERS. Psychedelic symptoms of cannabis and cocaine use as a function of trait impulsivity. Journal of Psychopharmacology. 2015, 29(3), 324-334. ISSN 0269-8811. Dostupné z: doi:10.1177/0269881114563633

[21] KUC, Joanna, Hannes KETTNER, Fernando ROSAS, David ERRITZOE, Eline HAIJEN, Mendel KAELEN, David NUTT a Robin L. CARHART-HARRIS. Psychedelic experience dose-dependently modulated by cannabis: results of a prospective online survey. Psychopharmacology. 2022, 239(5), 1425-1440. ISSN 0033-3158. Dostupné z: doi:10.1007/s00213-021-05999-1

[22] CARHART-HARRIS, Robin L a Guy M GOODWIN. The Therapeutic Potential of Psychedelic Drugs: Past, Present, and Future. Neuropsychopharmacology. 2017, 42(11), 2105-2113. ISSN 0893-133X. Dostupné z: doi:10.1038/npp.2017.84

[23] GARCIA-ROMEU, Albert a William A. RICHARDS. Current perspectives on psychedelic therapy: use of serotonergic hallucinogens in clinical interventions. International Review of Psychiatry. 2018, 30(4), 291-316. ISSN 0954-0261. Dostupné z: doi:10.1080/09540261.2018.1486289

[24] SANACORA, Gerard, Mark A. FRYE, William MCDONALD, Sanjay J. MATHEW, Mason S. TURNER, Alan F. SCHATZBERG, Paul SUMMERGRAD a Charles B. NEMEROFF. A Consensus Statement on the Use of Ketamine in the Treatment of Mood Disorders. JAMA Psychiatry. 2017, 74(4). ISSN 2168-622X. Dostupné z: doi:10.1001/jamapsychiatry.2017.0080

[25] ZANOS, Panos, Ruin MOADDEL, Patrick J. MORRIS, et al. Ketamine and Ketamine Metabolite Pharmacology: Insights into Therapeutic Mechanisms. Pharmacological Reviews. 2018, 70(3), 621-660. ISSN 0031-6997. Dostupné z: doi:10.1124/pr.117.015198

[26] BOGENSCHUTZ, Michael P, Alyssa A FORCEHIMES, Jessica A POMMY, Claire E WILCOX, PCR BARBOSA a Rick J STRASSMAN. Psilocybin-assisted treatment for alcohol dependence: A proof-of-concept study. Journal of Psychopharmacology. 2015, 29(3), 289-299. ISSN 0269-8811. Dostupné z: doi:10.1177/0269881114565144

[27] JOHNSON, Matthew W., Albert GARCIA-ROMEU a Roland R. GRIFFITHS. Long-term follow-up of psilocybin-facilitated smoking cessation. The American Journal of Drug and Alcohol Abuse. 2017, 43(1), 55-60. ISSN 0095-2990. Dostupné z: doi:10.3109/00952990.2016.1170135

[28] LABATE, Beatriz Caiuby a Henrik JUNGABERLE. The internationalization of ayahuasca. LIT Verlag Münster, 2011.

[29] NYBERG, Harri. Religious use of hallucinogenic fungi: A comparison between Siberian and Mesoamerican Cultures. Karstenia, 1992, 32.2: 71-80.

[30] FADIMAN, James. The psychedelic explorer's guide: Safe, therapeutic, and sacred journeys. Simon and Schuster, 2011.

[31] KUYPERS, Kim PC, Livia NG, David ERRITZOE, et al. Microdosing psychedelics: More questions than answers? An overview and suggestions for future research. Journal of Psychopharmacology. 2019, 33(9), 1039-1057. ISSN 0269-8811. Dostupné z: doi:10.1177/0269881119857204

[32] BASEDOW, Lukas A. a Sören KUITUNEN-PAUL. Motives for the use of serotonergic psychedelics: A systematic review. Drug and Alcohol Review. 2022, 41(6), 1391-1403. ISSN 0959-5236. Dostupné z: doi:10.1111/dar.13480

[33] CARHART-HARRIS, Robin L, Leor ROSEMAN, Eline HAIJEN, David ERRITZOE, Rosalind WATTS, Igor BRANCHI a Mendel KAELEN. Psychedelics and the essential importance of context. Journal of Psychopharmacology. 2018, 32(7), 725-731. ISSN 0269-8811. Dostupné z: doi:10.1177/0269881118754710

[34] LEARY, Timothy, Ralph METZNER a Ram DASS. The Psychedelic Experience: A Manual Based on the Tibetan Book of the Dead. University of Virginia: University Books, 1964. ISBN 9780806505527.

[35] DUŘT, Martin. Rizika psychedelik. Utheraptor [online]. [cit. 2023-03-16]. Dostupné z: https://utheraptor.art/2019/03/14/rizika-psychedelik/

[36] JOHNSON, MW, WA RICHARDS a RR GRIFFITHS. Human hallucinogen research: guidelines for safety. Journal of Psychopharmacology. 2008, 22(6), 603-620. ISSN 0269-8811. Dostupné z: doi:10.1177/0269881108093587

[37] DITTRICH, A. The Standardized Psychometric Assessment of Altered States of Consciousness (ASCs) in Humans. Pharmacopsychiatry. 1998, 31(S 2), 80-84. ISSN 0176-3679. Dostupné z: doi:10.1055/s-2007-979351

[38] STUDERUS, Erich, Alex GAMMA, Franz X. VOLLENWEIDER a Vaughan BELL. Psychometric Evaluation of the Altered States of Consciousness Rating Scale (OAV). PLoS ONE. 2010, 5(8). ISSN 1932-6203. Dostupné z: doi:10.1371/journal.pone.0012412

[39] SCHMIDT, Timo T. a Hendrik BERKEMEYER. The Altered States Database: Psychometric Data of Altered States of Consciousness. Frontiers in Psychology. 2018, 9. ISSN 1664-1078. Dostupné z: doi:10.3389/fpsyg.2018.01028

[40] NOUR, Matthew M., Lisa EVANS, David NUTT a Robin L. CARHART-HARRIS. Ego-Dissolution and Psychedelics: Validation of the Ego-Dissolution

Inventory (EDI). Frontiers in Human Neuroscience. 2016, 10. ISSN 1662-5161. Dostupné z: doi:10.3389/fnhum.2016.00269

[41] BARRETT, Frederick S, Matthew W JOHNSON a Roland R GRIFFITHS. Validation of the revised Mystical Experience Questionnaire in experimental sessions with psilocybin. Journal of Psychopharmacology. 2015, 29(11), 1182-1190. ISSN 0269-8811. Dostupné z: doi:10.1177/0269881115609019

[42] ROSEMAN, Leor, David J. NUTT a Robin L. CARHART-HARRIS. Quality of Acute Psychedelic Experience Predicts Therapeutic Efficacy of Psilocybin for Treatment-Resistant Depression. Frontiers in Pharmacology . 2018, 8. ISSN 1663-9812. Dostupné z: doi:10.3389/fphar.2017.00974

[43] KAŁUŻNA, Ada, Marco SCHLOSSER, Emily GULLIKSEN CRASTE, Jack STROUD a James COOKE. Being no one, being One: The role of ego-dissolution and connectedness in the therapeutic effects of psychedelic experience. Journal of Psychedelic Studies. 2022, 6(2), 111-136. ISSN 2559-9283. Dostupné z: doi:10.1556/2054.2022.00199

[44] GÉRON, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. Second edition. Beijing: O'Reilly, 2019. ISBN 978-1-492-03264-9.

[45] ALPAYDIN, Ethem. Introduction to machine learning. 2nd ed. Massachusetts: MIT Press, 2010. ISBN 978-0-262-01243-0.

[46] TAN, Pang-Ning, Michael STEINBACH a Vipin KUMAR. Introduction to data mining. Harlow: Pearson Education, 2013. Pearson new international edition. ISBN 978-1-292-02615-2.

[47] RUSSELL, Stuart J. a Peter NORVIG. Artificial intelligence: a modern approach. 3rd ed. Harlow: Pearson Education, 2014. ISBN 978-1-29202-420-2.

[48] CARDOSO, Jaime S. a Ricardo SOUSA. Measuring the Performance of Ordinal Classification. International Journal of Pattern Recognition and Ar-

tificial Intelligence. 2012, 25(08), 1173-1195. ISSN 0218-0014. Dostupné z: doi:10.1142/S0218001411009093

[49] PEŠOUT, Štěpán. Drug Consumption by Personality Data [online]. Universidade de Évora, 2022 [cit. 2023-03-18]. Dostupné z: https://pesout.net/projects/dcpd.pdf

[50] SHMUELI, Boaz. Multi-Class Metrics Made Simple, Part I: Precision and Recall. Towards Data Science [online]. 2019 [cit. 2023-03-18]. Dostupné z: https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

[51] KUMAR, Satyam. 7 Ways to Handle Missing Values in Machine Learning: Popular strategies to handle missing values in the dataset. Towards Data Science [online]. 2020 [cit. 2023-03-19]. Dostupné z: https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e

[52] XU-YING Liu, JIANXIN Wu a ZHI-HUA Zhou. Exploratory Undersampling for Class-Imbalance Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2009, 39(2), 539-550. ISSN 1083-4419. Dostupné z: doi:10.1109/TSMCB.2008.2007853

[53] SHUO Wang a XIN Yao. Using Class Imbalance Learning for Software Defect Prediction. IEEE Transactions on Reliability [online]. 2013, 62(2), 434-443 [cit. 2023-03-20]. ISSN 0018-9529. Dostupné z: doi:10.1109/TR.2013.2259203

[54] GOLINKO, Eric, Thomas SONDERMAN a Xingquan ZHU. CNFL: Categorical to Numerical Feature Learning for Clustering and Classification. In: 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC). IEEE, 2017, 2017, s. 585-594. ISBN 978-1-5386-1600-0. Dostupné z: doi:10.1109/DSC.2017.87

[55] BROWNLEE, Jason. How to use Data Scaling Improve Deep Learning Model Stability and Performance. Machine Learning Mastery [online]. 2019 [cit. 2023-03-19]. Dostupné z: https://machinelearningmastery.com/how-to-improve-neural-network-stability-and-modeling-performance-with-data-scaling/

[56] GUYON, Isabelle a André ELISSEEFF. An Introduction of Variable and Feature Selection. Journal of Machine Learning Research. 2003. Dostupné z: doi:10.1162/153244303322753616

[57] RASCHKA, Sebastian a Vahid MIRJALILI. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition. 2019. Packt Publishing. ISBN 978-1789955750.

[58] PRIYAM, Anuja, et al. Comparative analysis of decision tree classification algorithms. International Journal of current engineering and technology, 2013, 3.2: 334-337.

[59] KOTSIANTIS, S. B. Decision trees: a recent overview. Artificial Intelligence Review. 2013, 39(4), 261-283. ISSN 0269-2821. Dostupné z: doi:10.1007/s10462-011-9272-4

[60] BIAU, Gérard a Erwan SCORNET. A random forest guided tour. TEST. 2016, 25(2), 197-227. ISSN 1133-0686. Dostupné z: doi:10.1007/s11749-016-0481-7

[61] OSHIRO, Thais Mayumi, Pedro Santoro PEREZ a José Augusto BARANAUSKAS. How Many Trees in a Random Forest?. In: PERNER, Petra, ed. Machine Learning and Data Mining in Pattern Recognition. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, 2012, s. 154-168. Lecture Notes in Computer Science. ISBN 978-3-642-31536-7. Dostupné z: doi:10.1007/978-3-642-31537-4_13

[62] AO, Yile, Hongqi LI, Liping ZHU, Sikandar ALI a Zhongguo YANG. The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. Journal of Petroleum Science and Engineering. 2019, 174, 776-789. ISSN 09204105. Dostupné z: doi:10.1016/j.petrol.2018.11.067

[63] KOTSIANTIS, Sotiris B., et al. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 2007, 160.1: 3-24.

[64] ZHANG, Harry. Exploring Conditions for the Optimality of Naïve Bayes. International Journal of Pattern Recognition and Artificial Intelligence. 2011, 19(02),

183-198. ISSN 0218-0014. Dostupné z: doi:10.1142/S0218001405003983

[65] Naive Bayes. Scikit-learn [online]. 2023 [cit. 2023-03-21]. Dostupné z: https://scikit-learn.org/stable/modules/naive_bayes.html

[66] Jakkula, V. (2006). Tutorial on support vector machine (svm). School of EECS, Washington State University, 37(2.5), 3.

[67] Support vector machines (SVM): Algoritmy podpůrných vektorů [online]. Masarykova univerzita, 2006 [cit. 2023-03-22]. Dostupné z: https://is.muni.cz/el/1433/podzim2006/PA034/09_SVM.pdf

[68] KŘIVAN, Miloš. Umělé neuronové sítě. Praha: Oeconomica, 2021. ISBN 978-80-245-2420-7.

[69] HAKL, František a Martin HOLEŇA. Úvod do teorie neuronových sítí. Praha: ČVUT Praha a AV ČR, 1997. Dostupné z: http://cmp.felk.cvut.cz/cmp/courses/recognition/resources/_NN/hakl_nn98.pdf

[70] Neural Networks Concepts. ML Glossary [online]. 2017 [cit. 2023-03-23]. Dostupné z: https://ml-cheatsheet.readthedocs.io/en/latest/nn_concepts.html

[71] KOSTADINOV, Simeon. Understanding Backpropagation Algorithm. Towards Data Science [online]. 2019 [cit. 2023-03-23]. Dostupné z: https://towardsdatascience.com/understanding-backpropagation-algorithm-7bb3aa2f95fd

[72] VOSOL, David. Využití evolučních algoritmů při učení neuronových sítí. Brno, 2018. Bakalářská práce. VUT v Brně.

[73] BENI, Gerardo a WANG, Jing. Swarm intelligence in cellular robotic systems. Robots and biological systems: towards a new bionics?. Springer Berlin Heidelberg, 1993. p. 703-712.

[74] KARGER, Michal. Algoritmus Diferenciální Evoluce s prvky deterministického chaosu (ChaosDE) v prostředí Mathematica. Zlín, 2011. Dostupné z:

https://theses.cz/id/xqgoho/. Diplomová práce. Univerzita Tomáše Bati ve Zlíně, Fakulta aplikované informatiky.

[75] iTrip: Mobilní aplikace sbírá data, poskytuje informace a snižuje rizika spojená s užíváním psychedelických látek. ČTK [online]. Praha, 2021 [cit. 2023-03-25]. Dostupné z: https://www.ceskenoviny.cz/zpravy/itrip-mobilni-aplikace-sbira-data-poskytuje-informace-a-snizuje-rizika-spojena-s-uzivanim-psychedelickych-latek/2045709