VYSOKÁ ŠKOLA POLYTECHNICKÁ JIHLAVA

Aplikovaná informatika

DATOVÁ VĚDA: ANALÝZA STUDIJNÍ ÚSPĚŠNOSTI

Bakalářská práce

Autor práce: Pavel Černý

Vedoucí práce: doc. Dr. Ing. Jan Voráček, CSc.

Jihlava 2026

Vysoká škola polytechnická Jihlava

Tolstého 16, 586 01 Jihlava

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Autor práce:	**Pavel Černý**
Studijní program:	Aplikovaná informatika
Garant studijního programu:	Ing. Lenka Kuklišová Pavelková, Ph.D.

Název práce:	**Datová věda: analýza studijní úspěšnosti**
Vedoucí práce:	doc. Dr. Ing. Jan Voráček, CSc.
Cíl práce:	Cílem práce je aplikace úplného životního cyklu datové vědy na problematiku analýzy studijní úspěšnosti vysokoškolských studentů. Půjde při ní zejména o nalezení nejvýznamnějších faktorů, determinujících výsledky ve vybraných oblastech studia. Naplnění cíle zahrnuje identifikaci problému, rešerši souvisejících zdrojů, návrh a optimalizaci vhodných experimentů, čištění a konsolidaci relevantních dat, jejich explorativní analýzy a induktivní modelování, srozumitelnou vizualizaci výsledků a diskusi možných řešení definovaného problému na základě shromážděných, vyhodnocených a řádně diskutovaných podkladů. Základní datový korpus bude převzat z veřejných zdrojů a student ho může dle vlastních úvah dále modifikovat. Experimenty budou realizovány a distribuovány formou interaktivního online notebooku, využívajícího problémově orientované knihovny jazyka Python. Při vývoji finálního řešení bude využita agilní metodika Rapid Application Development s podporou umělé inteligence.

Abstrakt

Bakalářská práce aplikuje životní cyklus datové vědy na analýzu studijní úspěšnosti vysokoškoláků. Cílem je identifikovat faktory determinující výsledky a ověřit jejich využití pro predikci úspěšnosti pomocí algoritmů strojového učení. Teoretická část shrnuje principy datové vědy, rámec CRISP-DM a metody analýzy vzdělávacích dat. Praktická část využívá dataset OULAD. Proces zahrnuje čištění dat, explorativní analýzu a modelování algoritmem XGBoost s automatizovanou optimalizací hyperparametrů. Experimenty jsou realizovány v jazyce Python formou interaktivního notebooku. Při vývoji prototypu dashboardu byla aplikována agilní metodika Rapid Application Development. Výsledky identifikují klíčové prediktory úspěšnosti a potvrzují vysoký potenciál datové vědy v pedagogickém kontextu. Poznatky mohou sloužit jako podklad pro další výzkum v oblasti Learning Analytics a pro podporu rozhodování při návrhu studijních strategií a včasných intervenčních opatření. Zvolený přístup demonstruje efektivní propojení technické analýzy s praktickými potřebami akademických institucí.

Klíčová slova

datová věda; strojové učení; learning analytics; xgboost; prediktivní modelování; rapid application development

Abstract

This bachelor thesis applies the data science life cycle to analyze student academic success in higher education. The objective is to identify factors determining academic results and to evaluate their use for success prediction using machine learning algorithms. The theoretical part summarizes data science principles, the CRISP-DM framework, and methods for educational data analysis. The practical part utilizes the OULAD dataset. The process involves data cleaning, exploratory analysis, and modeling using the XGBoost algorithm with automated hyperparameter optimization. Experiments are implemented in Python as an interactive notebook. The agile Rapid Application Development methodology was applied during the development of a dashboard prototype. The results identify key predictors of success and confirm the high potential of data science in a pedagogical context. The findings can serve as a basis for further research in Learning Analytics and provide decision support for designing study strategies and early intervention measures. The chosen approach demonstrates an effective link between technical analysis and the practical needs of academic institutions.

Keywords

data science; machine learning; learning analytics; xgboost; predictive modeling; rapid application development

Prohlašuji, že předložená bakalářská práce je původní a zpracoval/a jsem ji samostatně. Prohlašuji, že citace použitých pramenů je úplná, že jsem v práci neporušil/a autorská práva (ve smyslu zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů, v platném znění, dále též „**AZ**“).

Byl/a jsem seznámen/a s tím, že na mou bakalářskou práci se plně vztahuje **AZ**, zejména § 60 (školní dílo).

Podle § 47b zákona o vysokých školách souhlasím se zveřejněním své práce podle Směrnice pro vedení, vypracování a zveřejňování závěrečných prací na VŠPJ, a to bez ohledu na výsledek obhajoby.

Beru na vědomí, že VŠPJ má právo na uzavření licenční smlouvy o užití mé bakalářské práce a prohlašuji, že **s o u h l a s í m** s případným užitím mé bakalářské práce (prodej, zapůjčení apod.).

Jsem si vědom/a toho, že užít své bakalářské práce či poskytnout licenci k jejímu využití mohu jen se souhlasem VŠPJ, která má právo ode mě požadovat přiměřený příspěvek na úhradu nákladů, vynaložených vysokou školou na vytvoření díla (až do jejich skutečné výše), z výdělku dosaženého v souvislosti s užitím díla či poskytnutím licence.

V Jihlavě dne 13. dubna 2026

.....

Podpis studenta/ky

Poděkování

Rád bych poděkoval svému vedoucímu práce, doc. Dr. Ing. Janu Voráčkovi, CSc., za odborné vedení, cenné rady a podporu během zpracování této bakalářské práce.

Obsah

Seznam obrázků **7**
Seznam zkratek **8**
Úvod **9**
1 Teorie datové vědy **10**
1.1 Historie a vývoj datové vědy 10
1.2 Principy datové vědy 11
1.3 Metody a techniky v analýze dat 12
1.4 Faktory ovlivňující studijní úspěšnost 12
1.5 Životní cyklus datově orientovaného projektu 13
1.6 Learning Analytics a Educational Data Mining 14
1.7 Predikční modelování ve vzdělávacích datech 15
1.8 Datové zdroje ve vzdělávání 16
1.9 Kvalita dat a její vliv na výsledky analýzy 17
1.10 Interpretovatelnost a vysvětlitelnost modelů 18
1.11 Hodnocení modelů a validační strategie 19
2 Metodika **20**
2.1 Implementační prostředí a knihovny 20
2.2 Zvolený výzkumný přístup a metodologický rámec (RAD) 21
2.3 Obecné fáze přípravy a zpracování dat 22
3 Výsledky analýz **33**
3.1 Srovnání modelů 33
3.2 Interpretace modelu (Feature Importance) 34
3.3 Matice záměn (Confusion Matrix) a limity aplikace 36
3.4 Omezení práce a diskuse 38
4 Závěrečné hodnocení **39**
4.1 Shrnutí hlavních přínosů vzhledem ke zkoumanému problému 39
4.2 Míra naplnění cílů a výzkumná omezení 40
4.3 Doporučení pro další výzkum 41
Závěr **42**
Seznam použité literatury **43**
Přílohy **45**

Seznam obrázků

Obr. 1: Importy knihoven 21
Obr. 2: Schéma propojení tabulek datasetu OULAD 24
Obr. 3: Konsolidace dat do proměnné sum_click 25
Obr. 4: Detekce odlehlých hodnot v aktivitě (sum_click) 26
Obr. 5: Korelační matice a distribuce aktivity 27
Obr. 6: Segmentace: Průměrná aktivita v modulech podle úspěšnosti 28
Obr. 7: Vizualizace shluků (K-Means Clustering) 29
Obr. 8: Implementace objektivní funkce pro Bayesovskou optimalizaci 31
Obr. 9: Tvorba UI rozhraní 32
Obr. 10: ROC křivka (Receiver Operating Characteristic) 34
Obr. 11: Důležitost příznaků u modelu XGBoost 35
Obr. 12: Matice záměn finálního modelu 36
Obr. 13: Precision-Recall Křivka 37
Obr. 14: Workflow diagram aplikace 39
Obr. 15: Uživatelské rozhraní pro hromadný import externích dat 39
Obr. 16: Rozhraní pro experimentální parametrizaci a interaktivní ladění modelu XGBoost 40
Obr. 17: Prediktivní dashboard pro individuální analýzu studijního rizika 40

Seznam zkratek

ABT	Analytical base table (analytická základní tabulka)
AI	Artificial intelligence (umělá inteligence)
AUC	Area under the curve (plocha pod křivkou)
CRISP-DM	Cross-industry standard process for data mining (standardizovaný proces analytického projektu)
CSV	Comma-separated values (formát souboru s hodnotami oddělenými čárkou)
EDA	Exploratory data analysis (explorativní analýza dat)
IMD	Index of multiple deprivation (index vícenásobného znevýhodnění)
LIME	Local interpretable model-agnostic explanations (technika vysvětlování modelů)
LMS	Learning management system (systém pro řízení výuky)
MAE	Mean absolute error (průměrná absolutní chyba)
MOOC	Massive open online course (masivní online kurz)
MSE	Mean squared error (střední kvadratická chyba)
NLP	Natural language processing (zpracování přirozeného jazyka)
OULAD	Open university learning analytics dataset
RAD	Rapid application development (rychlý vývoj aplikací)
ROC	Receiver operating characteristic (operační charakteristika přijímače)
SHAP	SHapley additive exPlanations (technika vysvětlování modelů)
UI	User interface (uživatelské rozhraní)
VLE	Virtual learning environment (virtuální studijní prostředí)
XAI	Explainable artificial intelligence (vysvětlitelná umělá inteligence)
XGBoost	eXtreme gradient boosting

Úvod

Problematika studijní úspěšnosti ve vysokoškolském prostředí představuje aktuální a společensky významné téma. V éře digitalizace vzdělávání generují univerzitní systémy rozsáhlé objemy dat, jejichž hloubková analýza může vysokým školám poskytnout kritické informace pro zefektivnění vzdělávacích procesů a personalizaci studijní podpory. Obor datové vědy prochází v současnosti dynamickým rozvojem a nabízí pokročilé nástroje pro prediktivní modelování, které umožňují identifikovat rizika akademického neúspěchu dříve, než se stanou nezvratnými.

Cílem této bakalářské práce je aplikace úplného životního cyklu datové vědy na problematiku analýzy a predikce studijní úspěšnosti vysokoškolských studentů. Hlavním záměrem je identifikace a interpretace nejvýznamnějších faktorů, které determinují studijní výsledky v online vzdělávacím prostředí. Práce se zaměřuje na návrh a realizaci komplexního analytického postupu, který zahrnuje identifikaci problému, čištění a konsolidaci dat a následnou explorativní analýzu vedoucí k induktivnímu modelování.

Pro účely experimentální části je využit veřejně dostupný dataset OULAD (Open University Learning Analytics Dataset). Praktické řešení je realizováno a distribuováno formou interaktivního online notebooku s využitím programovacího jazyka Python a jeho specializovaných knihoven pro strojové učení. Při vývoji finálního softwarového prototypu je aplikována agilní metodika Rapid Application Development (RAD) s integrací prvků umělé inteligence, což umožňuje rychlou iteraci a vytvoření funkčního nástroje pro podporu rozhodování studijních poradců.

1 Teorie datové vědy

Datová věda představuje moderní interdisciplinární oblast, která spojuje statistiku, informatiku, matematiku a doménové znalosti s cílem získávat z dat relevantní informace a podporovat rozhodování. Její význam spočívá nejen ve schopnosti analyzovat historická data, ale také v predikci budoucího chování a identifikaci klíčových faktorů ovlivňujících různé procesy. Vzdělávací prostředí, finance, zdravotnictví či doprava jsou jen některé z oblastí, kde metody datové vědy nacházejí široké uplatnění. Pochopení teoretických principů, metod a životního cyklu datově orientovaných projektů je proto nezbytným předpokladem pro efektivní aplikaci analytických nástrojů a interpretaci výsledků. Následující kapitoly postupně představují historický vývoj datové vědy, její principy, používané metody a techniky, faktory ovlivňující studijní úspěšnost a související koncepty v oblasti vzdělávání.

1.1 Historie a vývoj datové vědy

Počátky datové vědy se datují do období, kdy se začala formovat moderní statistika. Je to konkrétně v první polovině 20. století, kdy poprvé vznikaly metody na zpracování větších objemů dat. Cílem analytických procesů bylo odhalení souvislostí mezi daty a identifikace hlavních faktorů, které ovlivňovaly povahu dat. Prvotní metody pro zpracování byly striktně statistické a spjaté s teoretickým modelováním, přičemž důraz byl kladen zejména na inferenční statistiku a testování hypotéz (Press, 2013). Analýza dat byla v této fázi značně limitována výpočetními možnostmi tehdejší techniky a probíhala převážně manuálně nebo s využitím jednoduchých mechanických nástrojů.

Další významný rozvoj datově orientovaných přístupů nastával postupně s vývojem výkonnějších výpočetních technologií. Rozšíření elektronických počítačů ve druhé polovině 20. století umožnilo automatizaci výpočtů a práci s výrazně většími datovými soubory. V tomto období se začaly objevovat první databázové systémy a metody pro ukládání a správu dat, které položily základy pozdějším analytickým přístupům. Analýza dat se postupně přesouvala od čistě teoretických modelů k praktičtějším aplikacím v průmyslu a výzkumu.

Samotný termín „data science“ lze vysledovat již do roku 1974, nicméně do širšího povědomí odborné veřejnosti se dostal až v 90. letech minulého století. V této době byl používán především jako označení pro aplikaci statistických metod na rozsáhlejší datové soubory s využitím moderních výpočetních prostředků. Teprve na přelomu tisíciletí začal být pojem datové vědy chápán jako samostatná interdisciplinární disciplína, která propojuje statistiku, informatiku, databázové systémy a strojové učení (Dhar, 2013).

Významný impulz pro rozvoj datové vědy představovalo rozšíření internetu a následně sociálních sítí, které umožnily shromažďování masivního množství heterogenních dat. Vývoj vedl ke vzniku konceptu tzv. „big data“ a k potřebě nových analytických metod schopných pracovat s velkým objemem, rychlostí a rozmanitostí dat. Datová věda se tak začala orientovat nejen na analýzu historických dat, ale také na jejich využití v reálném čase.

Specifikum datové vědy spočívá v její schopnosti automatizovaně objevovat závislosti v datech a využívat je pro predikci a podporu rozhodování. Odlišuje se tak od tradiční statistiky, která je zaměřena především na popis a vysvětlení jevů (Fawcett & Provost, 2013). Datová věda dnes

zahrnuje široké spektrum disciplín, od matematiky a statistiky přes informatiku až po doménové znalosti konkrétních aplikačních oblastí.

V současnosti datová věda prochází dynamickým rozvojem, který je poháněn zejména rozvojem metod strojového učení a rostoucí dostupností datových zdrojů. Metody datové vědy se uplatňují v řadě oblastí, včetně zdravotnictví, financí, dopravy či vzdělávání. Vývoj oboru tak reflektuje přechod od tradiční statistické analýzy ke komplexnímu zpracování dat v rámci celého jejich životního cyklu (Cielen, Meysman & Ali, 2016).

1.2 Principy datové vědy

Datová věda představuje interdisciplinární obor, jehož hlavním cílem je získávání relevantních informací z dat a jejich následné využití pro rozhodování. Základním principem datové vědy je systematický přístup k práci s daty, který zahrnuje celý proces od formulace analytického problému až po interpretaci výsledků a jejich aplikaci v praxi. Na rozdíl od tradičních analytických přístupů se datová věda zaměřuje nejen na popis dat, ale především na objevování vzorů, predikci a induktivní modelování (Dhar, 2013).

Klíčovým principem je správná formulace analytického problému. Před samotnou analýzou je nezbytné porozumět doméně, ve které se problém nachází, a jasně definovat cíle a očekávané výstupy. Nesprávně formulovaný problém může vést k analyticky správným, avšak prakticky nevyužitelným výsledkům. V oblasti vzdělávání je proto důležité, aby analytické cíle odpovídaly pedagogickým a institucionálním potřebám.

Dalším zásadním principem je důraz na kvalitu dat a jejich přípravu. Proces získávání, čištění, konsolidace a transformace dat představuje časově nejnáročnější část analytického cyklu. Kvalita vstupních dat má přímý vliv na spolehlivost výsledků a chybně připravená data mohou vést k zavádějícím závěrům. Správný návrh datových struktur a systematická práce s daty umožňují efektivní analýzu a minimalizují riziko chyb (Kimball & Ross, 2013).

Z hlediska procesního řízení analytických projektů patří mezi základní principy datové vědy využívání osvědčených metodik. Jednou z nejrozšířenějších je CRISP-DM, která definuje šest fází analytického procesu a podporuje iterativní přístup k řešení problémů (Chapman et al., 2000). Iterativnost umožňuje průběžné zpřesňování analytického řešení na základě nových poznatků a výsledků.

Neméně důležitým principem je explorativní analýza dat, jejímž cílem je porozumění struktuře dat, identifikace anomálií a odhalení potenciálních vztahů mezi proměnnými. Explorativní přístupy hrají klíčovou roli zejména v počátečních fázích projektů a často ovlivňují volbu následných analytických metod (Cielen, Meysman & Ali, 2016).

Významným principem datové vědy je rovněž interpretovatelnost výsledků. I vysoce přesný model má omezenou hodnotu, pokud jeho výstupy nejsou srozumitelné pro koncové uživatele. V oblasti vzdělávání je proto kladen důraz na transparentnost analytických postupů a jasnou vizualizaci výsledků, které umožňují jejich praktické využití (Siemens & Long, 2011).

1.3 Metody a techniky v analýze dat

Analýza dat v rámci datové vědy využívá široké spektrum metod a technik, jejichž cílem je získat z dat smysluplné informace, identifikovat skryté vztahy a vytvářet prediktivní modely. Volba konkrétních metod závisí na charakteru dat, definovaném problému a požadované formě výstupů. V praxi se obvykle kombinuje více analytických přístupů odpovídajících jednotlivým fázím životního cyklu datové vědy (Chapman et al., 2000).

Základním krokem analytického procesu je explorační analýza dat (EDA). Jejím cílem je porozumět struktuře dat, identifikovat základní statistické vlastnosti a odhalit případné anomálie. EDA často zahrnuje popisné statistiky, vizualizaci rozdělení hodnot a analýzu vztahů mezi proměnnými. Výsledky explorativní analýzy slouží jako podklad pro rozhodování o dalším postupu a volbě vhodných modelů (Cielen, Meysman & Ali, 2016).

Klíčovým mezikrokem mezi přípravou dat a samotným modelováním je tzv. Feature Engineering (inženýrství příznaků). Proces zahrnuje transformaci surových dat do podoby, která lépe reprezentuje studovaný problém a zvyšuje predikční schopnost algoritmů. V prostředí Learning Analytics je transformace nezbytná zejména při zpracování transakčních logů z LMS systémů. Surové záznamy čítající miliony řádků o jednotlivých interakcích studenta (kliknutí na materiály, zhlédnutí videa) musí být agregovány do smysluplných ukazatelů, jako je například kumulativní aktivita nebo frekvence zapojení. Vytvořené metriky pak slouží jako kvantitativní vyjádření studijního úsilí, které má pro model mnohem vyšší informační hodnotu než izolované časové značky.

Další významnou skupinou metod jsou statistické techniky, které umožňují testování hypotéz a kvantifikaci vztahů mezi proměnnými. Patří sem například korelační analýza, regresní modely nebo analýza rozptylu. Metody jsou ceněny zejména pro svou interpretovatelnost a schopnost vysvětlit souvislosti v datech (Dhar, 2013).

S rozvojem výpočetní techniky a dostupností rozsáhlých datových souborů se stále více uplatňují metody strojového učení. Metody jsou zaměřeny především na klasifikaci a predikci a umožňují zachytit i složité nelineární vztahy v datech. Mezi často používané algoritmy patří rozhodovací stromy, logistická regrese, metody založené na ensemble přístupech nebo neuronové sítě (Fawcett & Provost, 2013).

Nedílnou součástí analytických postupů je také vizualizace dat, která umožňuje srozumitelnou prezentaci výsledků a podporuje jejich interpretaci. Vizualizace hraje klíčovou roli nejen při explorativní analýze, ale i při komunikaci výsledků směrem ke koncovým uživatelům (Siemens & Long, 2011).

1.4 Faktory ovlivňující studijní úspěšnost

Studijní úspěšnost vysokoškolských studentů je komplexní jev, který je ovlivňován celou řadou vzájemně provázaných faktorů. Výzkumy v oblasti learning analytics a educational data mining ukazují, že neexistuje jediný univerzální ukazatel, který by byl schopen studijní úspěšnost spolehlivě vysvětlit. Naopak se jedná o kombinaci akademických, behaviorálních a demografických charakteristik studentů (Kumar & Vijayan, 2020).

Mezi nejvýznamnější faktory patří předchozí akademické výsledky. Dosavadní studijní výkon, například výsledky z předchozího studia nebo průběžné hodnocení v aktuálním kurzu, bývá považován za jeden z nejsilnějších prediktorů budoucí úspěšnosti. Analýzy vzdělávacích dat opakovaně potvrzují, že studenti s dobrými předchozími výsledky mají vyšší pravděpodobnost úspěšného dokončení studia (Macfadyen & Dawson, 2010).

Další významnou skupinu tvoří behaviorální faktory, které souvisejí s aktivitou studenta během studia. Patří sem například docházka, odevzdávání úkolů, aktivita v e-learningových systémech nebo zapojení do diskusí. Behaviorální data poskytují cenné informace ohledně studijních návyků studentů a umožňují včasnou identifikaci rizikového chování (Siemens & Long, 2011).

Neméně důležité jsou demografické a socioekonomické faktory, mezi které lze zařadit věk, pohlaví, formu studia nebo pracovní vytížení studenta. Ačkoli faktory obvykle nemají tak silný vliv jako akademické či behaviorální ukazatele, mohou významně přispět k vysvětlení rozdílů ve studijní úspěšnosti jednotlivých skupin studentů (Kuzilek, Hlosta & Zdrahal, 2017).

Identifikace faktorů ovlivňujících studijní úspěšnost je klíčová nejen pro predikci výsledků, ale také pro návrh cílených intervencí a zlepšení vzdělávacího procesu. Datová věda v tomto kontextu poskytuje nástroje, které umožňují nejen popis minulého chování studentů, ale také podporu rozhodování směřující k vyšší efektivitě vysokoškolského vzdělávání.

1.5 Životní cyklus datově orientovaného projektu

Životní cyklus datově orientovaného projektu představuje strukturovaný rámec, který popisuje jednotlivé fáze práce s daty od počáteční formulace problému až po interpretaci a nasazení výsledků. Umožňuje tak systematické řízení analytických aktivit a zajišťuje, že výsledky analýzy odpovídají původně stanoveným cílům. Na rozdíl od tradičních jednorázových analytických úloh je datová věda charakteristická iterativním charakterem, kdy se jednotlivé fáze projektu opakovaně zpřesňují na základě získaných poznatků (Cielen, Meysman & Ali, 2016).

Jedním z nejrozšířenějších modelů popisujících životní cyklus datového projektu je metodika CRISP-DM. Metodika rozděluje analytický proces do šesti základních fází: porozumění problému, porozumění datům, příprava dat, modelování, vyhodnocení a nasazení výsledků. Klíčovým přínosem zmíněného přístupu je důraz na propojení analytických činností s praktickými cíli a možnost návratu k předchozím fázím v případě potřeby úprav nebo zpřesnění řešení (Chapman et al., 2000).

Fáze porozumění problému je zásadní pro úspěch celého projektu, neboť definuje cíle analýzy, očekávané výstupy a kritéria úspěšnosti. V kontextu vzdělávacích dat je důležité správně formulovat výzkumné otázky tak, aby výsledky analýzy byly nejen statisticky relevantní, ale také prakticky využitelné. Následující fáze porozumění datům se zaměřuje na seznámení se s dostupnými datovými zdroji, jejich strukturou, kvalitou a omezeními, což umožňuje realisticky posoudit možnosti dalšího zpracování.

Příprava dat tvoří jednu z nejnáročnějších částí životního cyklu. Zahrnuje procesy čištění dat, odstraňování nekonzistencí, práci s chybějícími hodnotami a transformaci dat do vhodné podoby pro modelování. Kvalita této fáze má zásadní vliv na spolehlivost výsledných modelů, přičemž nedostatečná příprava dat může vést k zavádějícím závěrům (Kimball & Ross, 2013).

Ve fázi modelování jsou aplikovány vybrané analytické a predikční metody s cílem odhalit vztahy v datech nebo vytvořit modely schopné predikovat budoucí chování. Výběr konkrétních algoritmů závisí na povaze problému, charakteru dat a požadavcích na interpretovatelnost výsledků. Následující fáze vyhodnocení slouží k ověření kvality modelů a posouzení, zda splňují stanovené cíle. Hodnocení zahrnuje nejen měření přesnosti modelu, ale také analýzu jeho použitelnosti v reálném kontextu (Fawcett & Provost, 2013).

Závěrečnou fází životního cyklu je nasazení výsledků, které může mít podobu implementace modelu do informačního systému, vytvoření analytické zprávy nebo návrhu podpůrných opatření. V oblasti learning analytics je důležité, aby výsledky analýzy byly prezentovány srozumitelnou formou a byly využitelné pro rozhodování pedagogů či managementu vzdělávacích institucí (Siemens & Long, 2011).

V moderní datové vědě dochází k posunu od generování statických analytických zpráv směrem k vytváření interaktivních nástrojů pro podporu rozhodování. Tradiční cyklus CRISP-DM se proto v praxi stále častěji propojuje s agilními metodami vývoje softwaru, zejména s metodikou Rapid Application Development (RAD). Přístup klade důraz na rychlou tvorbu funkčních prototypů a iterativní vylepšování na základě zpětné vazby. V současnosti je proces navíc akcelerován přístupem AI-assisted development, kdy analytici využívají generativní umělou inteligenci k optimalizaci kódu, návrhu uživatelských rozhraní (dashboardů) a rychlejšímu překonávání technických bariér při implementaci modelů do produkčního prostředí.

Životní cyklus datově orientovaného projektu tak vytváří ucelený rámec, který propojuje technické aspekty analýzy dat s praktickými cíli a podporuje systematický přístup k řešení komplexních problémů, jako je analýza studijní úspěšnosti vysokoškolských studentů.

1.6 Learning Analytics a Educational Data Mining

Learning analytics a educational data mining představují specializované oblasti datové vědy zaměřené na analýzu dat vznikajících ve vzdělávacím prostředí. Jejich hlavním cílem je porozumět procesu učení, identifikovat faktory ovlivňující studijní výsledky a podpořit rozhodování vedoucí ke zlepšení vzdělávacího procesu. Rozvoj těchto oblastí úzce souvisí s rostoucí digitalizací vzdělávání a rozšířením e-learningových systémů, které generují velké množství dat o chování studentů (Siemens & Long, 2011).

Learning analytics se zaměřují především na sběr, analýzu a interpretaci dat s cílem porozumět a optimalizovat učení a prostředí, ve kterém probíhá. Důraz je kladen na propojení analytických výstupů s pedagogickou praxí a na podporu rozhodování učitelů, studentů i institucí. Významnou roli zde hraje vizualizace dat a srozumitelná prezentace výsledků, které umožňují rychlou identifikaci problémů a trendů ve studijním chování (Siemens & Long, 2011).

Educational data mining se naproti tomu soustředí více na technické aspekty analýzy dat a aplikaci algoritmů strojového učení a statistických metod. Cílem je automatizované objevování vzorů a vztahů ve vzdělávacích datech, které mohou sloužit například k predikci studijní úspěšnosti, detekci rizikových studentů nebo analýze efektivity výukových materiálů (Kumar & Vijayan, 2020). Educational data mining klade větší důraz na přesnost modelů a optimalizaci analytických postupů.

Přestože jsou learning analytics a educational data mining často vnímány jako samostatné oblasti, v praxi se výrazně prolínají. Obě disciplíny využívají podobné datové zdroje, jako jsou záznamy z learning management systémů, výsledky testů, demografické údaje studentů nebo informace o jejich aktivitě během studia. Rozdíl spočívá především v cílové skupině a způsobu interpretace výsledků, kdy learning analytics kladou větší důraz na praktickou využitelnost výstupů, zatímco educational data mining se více zaměřují na metodologickou stránku analýzy (Macfadyen & Dawson, 2010).

Významným přínosem obou přístupů je možnost včasné identifikace studentů ohrožených studijním neúspěchem. Analýza behaviorálních dat, jako je frekvence přihlášení do systému, aktivita při plnění úkolů nebo interakce s výukovými materiály, umožňuje odhalit vzorce chování, které mohou signalizovat zvýšené riziko neúspěchu. Poznatky lze následně využít k návrhu cílených intervencí a podpůrných opatření (Kuzilek, Hlosta & Zdrahal, 2017).

Využití learning analytics a educational data mining však přináší také určitá omezení a výzvy. Patří mezi ně zejména kvalita a úplnost dostupných dat, riziko zkreslení analytických modelů a nutnost odpovědné interpretace výsledků. Důležitým aspektem je rovněž ochrana osobních údajů a etické zacházení se studijními daty, což klade zvýšené nároky na transparentnost analytických postupů a komunikaci výsledků (Siemens & Long, 2011; Kumar & Vijayan, 2020).

Learning analytics a educational data mining tak představují klíčový teoretický rámec pro analýzu studijní úspěšnosti vysokoškolských studentů. Kombinací analytických metod datové vědy a doménových znalostí vzdělávacího procesu umožňují nejen zpětné hodnocení studijního chování, ale také podporu predikce a rozhodování směřující ke zlepšení kvality vysokoškolského vzdělávání.

1.7 Predikční modelování ve vzdělávacích datech

Predikční modelování představuje klíčovou součást analytických přístupů v oblasti datové vědy a hraje významnou roli také při analýze vzdělávacích dat. Jeho cílem je na základě historických dat vytvářet modely schopné odhadovat budoucí chování nebo výsledky, například pravděpodobnost úspěšného dokončení studia či riziko studijního neúspěchu. V kontextu vysokoškolského vzdělávání umožňuje predikční modelování včasnou identifikaci studentů, kteří mohou vyžadovat zvýšenou podporu (Dhar, 2013).

Základním předpokladem úspěšného predikčního modelování je vhodná volba vstupních proměnných. Ve vzdělávacích datech se často využívají kombinace akademických, behaviorálních a demografických faktorů. Správný výběr proměnných má zásadní vliv na kvalitu výsledných modelů, přičemž nadměrné množství nerelevantních atributů může vést ke snížení predikční schopnosti a zhoršení interpretovatelnosti modelu (Fawcett & Provost, 2013).

Ve vzdělávacích aplikacích jsou běžně využívány různé typy predikčních modelů, včetně regresních metod, klasifikačních algoritmů a modelů strojového učení. Mezi často používané patří logistická regrese, rozhodovací stromy nebo metody založené na ensemble přístupech. Metody umožňují zachytit jak lineární, tak nelineární vztahy v datech a přizpůsobit se specifickým charakteristikám analyzovaného datasetu (Dhar, 2013; Fawcett & Provost, 2013).

V rámci zmíněných ensemble přístupů se v posledních letech jako dominantní technika ukázal Gradient Boosting, a to zejména jeho vysoce optimalizovaná implementace v podobě algoritmu

XGBoost. Metoda funguje na principu sekvenčního budování rozhodovacích stromů, kde se každý nový strom učí z chyb těch předchozích. V oblasti Learning Analytics je technika ceněna pro svou mimořádnou přesnost, schopnost automaticky zpracovávat chybějící hodnoty a robustnost vůči odlehlým záznamům, což z ní činí jeden z nejvýkonnějších nástrojů pro predikci studijní úspěšnosti.

Důležitým aspektem predikčního modelování ve vzdělávání je rovnováha mezi přesností a interpretovatelností modelu. Zatímco složitější modely mohou dosahovat vyšší predikční úspěšnosti, jejich výstupy bývají obtížně srozumitelné pro pedagogy a další zainteresované osoby. Naopak jednodušší modely často umožňují lépe vysvětlit vliv jednotlivých faktorů na studijní úspěšnost, což je klíčové pro praktické využití výsledků (Siemens & Long, 2011).

Součástí procesu predikčního modelování je také důkladné vyhodnocení modelů. Kromě standardních metrik, jako je přesnost nebo míra chybovosti, je nutné posuzovat stabilitu modelu a jeho schopnost zobecnění na nová data. Ve vzdělávacím kontextu je rovněž důležité zohlednit časový aspekt dat a ověřit, zda model zůstává validní i při změnách studijních podmínek nebo struktury kurzu (Macfadyen & Dawson, 2010).

Predikční modelování ve vzdělávacích datech nachází uplatnění zejména při návrhu systémů včasného varování, které umožňují identifikovat rizikové studenty již v průběhu semestru. Systémy mohou sloužit jako podpora rozhodování pro pedagogy a studijní poradce a přispívat ke zvýšení studijní úspěšnosti prostřednictvím cílených intervencí (Kumar & Vijayan, 2020).

Predikční modelování tak představuje důležitý nástroj pro analýzu studijní úspěšnosti vysokoškolských studentů. Při jeho aplikaci je však nezbytné dbát na správnou interpretaci výsledků, etické aspekty a odpovědné využití analytických výstupů, aby získané poznatky skutečně přispívaly ke zlepšení vzdělávacího procesu.

1.8 Datové zdroje ve vzdělávání

Datové zdroje ve vzdělávání představují základní východisko pro aplikaci metod datové vědy na problematiku studijní úspěšnosti. S rozvojem digitalizace vzdělávacích procesů dochází k systematickému shromažďování velkého množství dat, která popisují studijní chování, výsledky i charakteristiky studentů. Data tvoří klíčový vstup pro oblasti educational data mining a learning analytics, jejichž cílem je porozumění vzdělávacím procesům a jejich optimalizace (Siemens & Long, 2011).

Jedním z nejčastějších zdrojů vzdělávacích dat jsou informační systémy vysokých škol, zejména studijní a administrativní systémy. Systémy obsahují strukturovaná data o studentech, jako jsou záznamy o zápisech do předmětů, průběžné a závěrečné hodnocení, studijní průměr, délka studia nebo informace o přerušení či ukončení studia. Administrativní data jsou relativně snadno analyzovatelná a často tvoří základ predikčních modelů studijní úspěšnosti, přestože samy o sobě neposkytují detailní vhled do studijního chování studentů.

Dalším významným zdrojem jsou data generovaná v systémech pro řízení výuky (Learning Management Systems, LMS), jako jsou Moodle, Blackboard nebo Canvas. Systémy zaznamenávají detailní informace o aktivitě studentů, například přihlašování do systému, prohlížení studijních materiálů, odevzdávání úkolů, výsledky testů či účast v diskusních fórech. Behaviorální data z LMS jsou považována za velmi cenná, protože umožňují analyzovat studijní

návyky studentů v průběhu semestru a včas identifikovat rizikové chování (Macfadyen & Dawson, 2010).

Specifickou kategorii představují data pocházející z externích vzdělávacích platforem a online kurzů, například MOOC systémů. Data mají často velký objem a vysokou granularitu, což umožňuje detailní analýzu interakcí studentů s výukovým obsahem. Zároveň však mohou být zatížena vyšší mírou neúplnosti nebo heterogenity, což klade zvýšené nároky na jejich předzpracování a interpretaci.

Z hlediska struktury lze vzdělávací data rozdělit na numerická, kategoriální a ordinální. Numerická data zahrnují například bodové hodnocení, studijní průměr nebo počet odevzdaných úkolů, zatímco kategoriální data mohou reprezentovat formu studia, typ předmětu nebo pohlaví studenta. Častým rysem vzdělávacích dat je také jejich časová povaha, kdy jsou jednotlivé záznamy vázány na konkrétní časový okamžik nebo studijní období. Skutečnost umožňuje sledovat vývoj studijního chování v čase, ale zároveň komplikuje analytické zpracování.

Při práci s datovými zdroji ve vzdělávání je nutné zohlednit jejich omezení a potenciální zkreslení. Data mohou být neúplná, zatížena chybami nebo nemusí přesně odrážet skutečné studijní úsilí studentů. Například vysoká aktivita v LMS nemusí nutně znamenat efektivní učení, a naopak nízká aktivita nemusí vždy indikovat studijní neúspěch. Interpretace výsledků analýzy proto vyžaduje znalost vzdělávací domény a kontextu, ve kterém data vznikla (Kuzilek, Hlosta & Zdrahal, 2017).

Datové zdroje ve vzdělávání tak představují nezbytný, avšak zároveň komplexní základ pro analytické a predikční úlohy. Jejich správná charakterizace, pochopení původu a struktury a uvědomění si jejich limitů jsou klíčovým předpokladem pro smysluplné využití metod datové vědy v oblasti hodnocení a predikce studijní úspěšnosti.

1.9 Kvalita dat a její vliv na výsledky analýzy

Kvalita dat představuje jeden z klíčových faktorů určujících spolehlivost, přesnost a praktickou využitelnost analytických výstupů v datové vědě. I nejpokročilejší modely a algoritmy strojového učení mohou generovat zavádějící či nepoužitelné výsledky, pokud vstupní data obsahují chyby, nesrovnalosti nebo neúplné informace (Cielen, Meysman & Ali, 2016). Kvalita dat se proto stává kritickým předpokladem úspěšné aplikace metod datové vědy, zejména v kontextu vzdělávacích dat, kde přesnost predikce studijní úspěšnosti přímo ovlivňuje rozhodování pedagogů a intervenční opatření (Kumar & Vijayan, 2020).

Mezi nejčastější problémy ovlivňující kvalitu dat patří chybějící hodnoty, nesprávně zaznamenané informace, duplicity, nekonzistence či odlehlé hodnoty, které mohou výrazně zkreslit výsledky analýzy (Kimball & Ross, 2013). Chybějící hodnoty například mohou vést k selektivnímu zkreslení, pokud jsou odstraněny nevhodným způsobem, nebo mohou snížit predikční schopnost modelů, pokud nejsou adekvátně doplněny. Odlehlé hodnoty mohou způsobit, že modely budou přizpůsobeny extrémním hodnotám, což negativně ovlivní jejich generalizační schopnost a interpretovatelnost (Fawcett & Provost, 2013).

Důležitou součástí zajištění kvality dat je také standardizace a validace zdrojů dat. Ve vzdělávacím prostředí se jedná například o porovnání záznamů z administrativních systémů s daty generovanými LMS, ověření konzistence výsledků testů a hodnocení studentů či

identifikaci chyb v časových řadách aktivit studentů (Macfadyen & Dawson, 2010; Kuzilek, Hlosta & Zdrahal, 2017). Nesoulad mezi jednotlivými datovými zdroji může vést k chybné interpretaci vztahů mezi proměnnými a ke zkreslení predikčních modelů.

Kvalita dat rovněž ovlivňuje volbu analytických metod. Například algoritmy strojového učení jsou citlivé na rozsah chyb a odlehlých hodnot, a proto je často nutné provádět normalizaci, standardizaci nebo transformace dat (Dhar, 2013). Pro klasifikační úlohy, jako je predikce úspěšnosti studentů, mohou nesprávně připravená data vést k falešně pozitivním či falešně negativním predikcím, což má přímý dopad na intervence a doporučení pedagogů.

Iterativní životní cyklus datového projektu, například definovaný metodikou CRISP-DM, klade důraz na opakované ověřování kvality dat v průběhu celého analytického procesu (Chapman et al., 2000). Zahrnuje tak nejen kontrolu vstupních dat, ale i průběžnou evaluaci modelů a predikcí s cílem odhalit případné chyby, nesrovnalosti nebo nové odlehlé hodnoty, které se mohou objevit při aktualizaci dat. V praxi to znamená, že kvalita dat není statická vlastnost, ale kontinuální proces, který vyžaduje pečlivou správu a monitoring.

V oblasti learning analytics a educational data mining má kvalita dat zvláštní význam. Behaviorální data, například z LMS, mohou být neúplná nebo zkreslená, pokud studenti nevyužívají systém konzistentně, či pokud některé interakce nejsou zaznamenány. Demografická a administrativní data mohou obsahovat chybné nebo zastaralé informace. Bez náležitého zpracování a validace by modely mohly vytvářet nesprávné predikce a vést k nesprávným závěrům o studijní úspěšnosti jednotlivých studentů (Siemens & Long, 2011).

Zajištění vysoké kvality dat proto zahrnuje kombinaci technických a procesních opatření – od čištění a transformace dat, přes kontrolu jejich konzistence a úplnosti, až po průběžné sledování integrity dat v reálném čase. Správně připravená a validovaná data jsou předpokladem nejen přesných predikcí, ale také srozumitelných a interpretovatelných výstupů, které mohou být efektivně využity pro podporu rozhodování a optimalizaci vzdělávacího procesu (Press, 2013; Fawcett & Provost, 2013).

Kvalita dat tedy není pouze technickou záležitostí, ale klíčovým determinantem spolehlivosti a použitelnosti výsledků datové analýzy. Bez jejího zajištění nelze očekávat, že analytické modely poskytnou validní informace, ať už se jedná o identifikaci faktorů ovlivňujících studijní úspěšnost, predikci výsledků studentů či návrh cílených intervencí.

1.10 Interpretovatelnost a vysvětlitelnost modelů

Interpretovatelnost modelů je jedním z klíčových principů datové vědy, zejména v oblastech, kde výsledky analýzy mají přímý dopad na rozhodování lidí, jako je vzdělávání (Dhar, 2013). I vysoce přesné modely strojového učení mohou být bez srozumitelného vysvětlení pro pedagogy a správce vzdělávacích institucí prakticky nepoužitelné. Interpretovatelnost zajišťuje, že uživatelé mohou pochopit, proč model dospěl k určité predikci, které proměnné měly největší vliv a jaké vztahy byly v datech identifikovány (Fawcett & Provost, 2013).

Ve vzdělávacím kontextu je interpretovatelnost kritická například při identifikaci rizikových studentů. Modely, které dokážou vysvětlit vliv jednotlivých akademických, behaviorálních či demografických faktorů na výsledky studenta, umožňují pedagogům cíleně zasahovat a poskytovat podporu tam, kde je nejvíce potřebná (Siemens & Long, 2011). Přitom platí, že

existuje kompromis mezi predikční přesností a vysvětlitelností: složité modely, jako jsou hluboké neuronové sítě nebo ensemble metody, mohou dosahovat vysoké přesnosti, avšak jejich vnitřní logika je často pro lidského uživatele obtížně srozumitelná (Cielen, Meysman & Ali, 2016).

Pro zvýšení interpretovatelnosti se často používají techniky vysvětlování modelů, například důležitost proměnných (feature importance), parcimonické modely (parsimonious model), rozhodovací stromy, LIME nebo SHAP, které poskytují post-hoc vysvětlení i pro složité modely (Fawcett & Provost, 2013). Transparentní vizualizace, jako jsou heatmapy korelací nebo parcely vlivu jednotlivých proměnných, rovněž napomáhají lepšímu pochopení výsledků. Všechny postupy jsou nezbytné pro to, aby výstupy datové vědy byly prakticky využitelné a aby modely podporovaly kvalifikovaná rozhodnutí pedagogů a administrativy (Kumar & Vijayan, 2020).

1.11 Hodnocení modelů a validační strategie

Hodnocení modelů je nezbytnou fází životního cyklu datově orientovaného projektu a určuje spolehlivost, přesnost a schopnost modelu generalizovat na nová data (Chapman et al., 2000). Správně navržené validační strategie umožňují odhalit přeučení (overfitting) nebo podučení (underfitting) a poskytují objektivní měřítko predikční schopnosti modelů, což je klíčové při aplikaci predikčních metod na vzdělávací data (Dhar, 2013).

Pro hodnocení modelů klasifikace se běžně používají metriky, jako je přesnost (accuracy), dosažitelnost (recall), preciznost (precision), F1-skóre nebo ROC AUC. Metriky poskytují komplexní pohled na výkon modelu a umožňují identifikovat případné slabiny v predikci jednotlivých tříd (Fawcett & Provost, 2013). U regresních modelů, například při predikci studijního průměru, se využívají metriky jako střední kvadratická chyba (MSE), průměrná absolutní chyba (MAE) nebo koeficient determinace R^2 (Cielen, Meysman & Ali, 2016).

Validace modelů se provádí různými způsoby. Tradiční přístup zahrnuje rozdělení datasetu na trénovací, validační a testovací sadu, například v poměru 60/20/20, což umožňuje odděleně trénovat model, ladit hyperparametry a posuzovat jeho schopnost generalizace (Kimball & Ross, 2013). Další často používanou metodou je křížová validace (cross-validation), která poskytuje robustnější odhad výkonu modelu zejména u menších datových souborů (Fawcett & Provost, 2013).

V oblasti vzdělávání je zvláštní pozornost věnována stabilitě modelu v čase. Data studentů mohou vykazovat sezónní či kurikulum-specifické variace, které je nutné zohlednit při hodnocení modelu. Validace musí zajistit, že model bude spolehlivě predikovat i při změně struktury kurzu nebo studentů (Macfadyen & Dawson, 2010).

Hodnocení modelů tedy nejen potvrzuje jejich predikční schopnost, ale zároveň poskytuje návod pro optimalizaci a výběr nejvhodnějšího algoritmu. Efektivní validační strategie je nezbytná pro zajištění toho, aby analytické výstupy byly spolehlivé, interpretovatelné a prakticky využitelné pro podporu rozhodování a intervencí ve vzdělávacím procesu (Siemens & Long, 2011; Kumar & Vijayan, 2020).

2 Metodika

Metodologický rámec této práce je navržen jako syntéza exaktních statistických metod a moderních agilních postupů softwarového vývoje. Cílem zvolené metodiky je nejen vytvořit přesný prediktivní model studijní úspěšnosti, ale také zajistit jeho praktickou aplikovatelnost v reálném akademickém prostředí. Celý proces transformace surových dat v relevantní znalosti a funkční prototyp se opírá o robustní procesní model CRISP-DM, který zajišťuje logickou návaznost jednotlivých kroků – od počátečního pochopení komplexity OULAD datasetu přes precizní inženýring příznaků (feature engineering) až po finální evaluaci. Vzhledem k dynamické povaze vývoje interaktivního nástroje je rámec doplněn o principy metodiky RAD (Rapid Application Development), která umožňuje cyklické zdokonalování modelu na základě okamžité vizuální zpětné vazby. Následující podkapitoly detailně popisují technické a analytické operace, které tvoří páteř praktické části práce a zaručují validitu i reprodukovatelnost dosažených výsledků.

2.1 Implementační prostředí a knihovny

Jádro celého analytického systému je postaveno na programovacím jazyce Python ve verzi 3.12, který byl zvolen pro svou dominanci v oblasti strojového učení a bohatý ekosystém specializovaných knihoven. Vývoj probíhal v interaktivním prostředí Kaggle Notebook, které umožňuje úzké propojení spustitelného kódu, vizualizací a textového popisu. Přístup je v souladu s principy reprodukovatelného výzkumu, neboť dovoluje krok po kroku dokumentovat proces transformace dat i ladění modelů.

Pro základní manipulaci s daty a inženýrství příznaků (feature engineering) byly klíčové knihovny Pandas a NumPy. Pandas poskytuje robustní datové struktury, zejména DataFrame, který je nezbytný pro efektivní spojování (join) heterogenních tabulek datasetu OULAD. NumPy doplňuje toto řešení o optimalizované matematické operace s poli, které tvoří nízkoúrovňový základ pro většinu algoritmů strojového učení. Nástroje umožnily transformovat miliony záznamů o interakcích studentů do agregované podoby vhodné pro trénování modelů.

V oblasti samotného modelování byla využita knihovna Scikit-learn, která představuje průmyslový standard pro klasické strojové učení. Byla nasazena především ve fázi předzpracování dat (preprocessing), konkrétně pro normalizaci příznaků a kódování kategoriálních proměnných. Dále sloužila k implementaci baseline modelů, jako je logistická regrese a Random Forest, které vytvořily srovnávací rámec pro finální řešení. Scikit-learn rovněž poskytla nástroje pro evaluaci modelů, včetně výpočtu metrik AUC, Precision a Recall.

Jako finální klasifikátor byl zvolen algoritmus XGBoost (Extreme Gradient Boosting). Knihovna implementuje metodu gradientního zvyšování stromů, která v současnosti dominuje v soutěžích zaměřených na predikci z tabulkových dat. XGBoost byl vybrán nejen pro svou vysokou predikční přesnost a efektivitu paralelního učení, ale také pro svou nativní schopnost pracovat s chybějícími hodnotami bez nutnosti složité manuální imputace. Algoritmus využívá pokročilé techniky regularizace, čímž efektivně předchází přetrénování modelu na trénovacích datech.

Proces optimalizace hyperparametrů byl automatizován pomocí moderního frameworku Optuna. Na rozdíl od tradičních metod, jako je GridSearch, využívá Optuna bayesovskou

optimalizaci a algoritmus TPE (Tree-structured Parzen Estimator), čímž umožňuje inteligentní prohledávání prostoru parametrů, kdy se algoritmus učí z předchozích neúspěšných pokusů a zaměřuje se na slibné oblasti, což vede k nalezení optimální konfigurace modelu v řádově kratším čase.

Uživatelské rozhraní a interaktivní prvky prototypu byly realizovány kombinací knihoven Plotly a Ipywidgets. Plotly umožňuje generování dynamických grafů, které na rozdíl od statických obrázků dovolují uživateli zkoumat jednotlivé datové body a trendy interaktivně. Ipywidgets pak tvoří ovládací vrstvu (tlačítka, posuvníky, nahrávací boxy), která přemosťuje složitý backend modelu s uživatelsky přívětivým prostředím. Integrace umožňuje studijním poradcům nahrávat aktuální data a okamžitě vizualizovat predikce bez nutnosti jakéhokoliv zásahu do zdrojového kódu.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import optuna
import warnings
import io
import base64
import ipywidgets as widgets
from IPython.display import display, clear_output
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from xgboost import XGBClassifier
from sklearn.metrics import (classification_report, roc_auc_score, confusion_matrix,
                             ConfusionMatrixDisplay, roc_curve, auc)
```

Obr. 1: Importy knihoven
Zdroj: vlastní zpracování

2.2 Zvolený výzkumný přístup a metodologický rámec (RAD)

Předkládaná práce kombinuje analytický přístup zaměřený na dolování znalostí z databází s principy moderního softwarového inženýrství. Pro celkové řízení datového projektu byl využit mezinárodně uznávaný průmyslový standard CRISP-DM (Cross-Industry Standard Process for Data Mining). Metodika definuje rigidní, ale zároveň cyklický životní cyklus projektu, který začíná hloubkovým pochopením obchodního či akademického problému, pokračuje přes náročnou přípravu dat a induktivní modelování až po finální evaluaci a nasazení výsledků do praxe. V kontextu této práce umožnil CRISP-DM udržet jasnou strukturu při zpracování rozsáhlých a heterogenních datových sad, aniž by došlo ke ztrátě fokusu na primární cíl – predikci studijní úspěšnosti.

Pro samotnou implementaci interaktivního analytického nástroje byla zvolena metodika RAD (Rapid Application Development). Agilní přístup klade důraz na rychlé vytváření funkčních prototypů a iterativní vývoj založený na průběžném testování a zpětné vazbě. V prostředí datové vědy se metodika RAD ukázala jako vysoce efektivní, neboť umožnila plynulý přechod od surových datových skriptů v jazyce Python k plně interaktivnímu uživatelskému rozhraní realizovanému prostřednictvím knihovny ipywidgets. Díky tomuto přístupu mohl být prototyp

okamžitě testován z pohledu koncového uživatele, například studijního poradce, což výrazně zkrátilo vývojový cyklus a umožnilo včasnou identifikaci potřebných úprav v layoutu a funkcionalitě dashboardu.

Vývoj finálního řešení byl dále akcelerován využitím nástrojů generativní umělé inteligence, což definuje moderní koncept AI-assisted development. Umělá inteligence hrála klíčovou roli především v iterativních fázích psaní a optimalizace kódu. Významně přispěla k refaktorizaci složitých transformací nad rozsáhlým datasetem studentVle, čítajícím přes 10 milionů záznamů, kde bylo nutné zajistit paměťovou efektivitu výpočtů. AI byla rovněž využita při návrhu uživatelského rozhraní, konkrétně při implementaci asynchronních prvků a komplexních layoutů v prostředí Kaggle Notebook. V neposlední řadě pomohla při definici objektivních funkcí pro knihovnu Optuna, což znatelně zkrátilo čas potřebný k nalezení optimální konfigurace hyperparametrů modelu XGBoost a vedlo k dosažení vyšší preciznosti výsledných predikcí.

2.3 Obecné fáze přípravy a zpracování dat

Fáze přípravy a zpracování dat představuje nejrozsáhlejší a technologicky nejnáročnější část celého analytického procesu. V kontextu predikce studijní úspěšnosti je kvalita výstupního modelu přímo závislá na reprezentativnosti a čistotě vstupních proměnných, neboť i sebesložitější algoritmus strojového učení poskytuje zavádějící výsledky, pokud jsou vstupní data nekvalitní (princip „Garbage In, Garbage Out“).

Proces zpracování v této práci zahrnuje transformaci surových, transakčních záznamů o interakcích studentů z tabulky studentVle do podoby strukturovaných statistických ukazatelů, které jsou následně integrovány s demografickými profily z tabulky studentInfo. Etapa vyžaduje hluboké porozumění sémantice jednotlivých atributů a aplikaci pokročilých technik čištění, jako je normalizace měřítek, imputace chybějících hodnot či kódování kategoriálních proměnných. Cílem je vytvoření tzv. „Analytical Base Table“ (ABT), která slouží jako jednotný a konzistentní podklad pro trénování klasifikačních algoritmů a následnou interpretaci klíčových faktorů úspěchu.

2.3.1 Charakteristika a zdroj použitých dat

Hlavním datovým zdrojem pro tuto práci je Open University Learning Analytics Dataset (OULAD), který je považován za jeden z nejkomplexnějších veřejně dostupných souborů dat v oblasti studijní analytiky. Dataset byl uvolněn britskou Open University (OU) s cílem podpořit výzkum v oblasti predikce studijních výsledků a prevence předčasného ukončení studia. OU je specifická svým zaměřením na distanční vzdělávání, což znamená, že interakce studenta s výukovými materiály probíhá téměř výhradně v digitálním prostředí, což umožňuje velmi přesné monitorování studijního úsilí.

Dataset jako celek se skládá ze sedmi vzájemně propojených relačních tabulek, které zachycují demografii studentů, jejich studijní historii, výsledky hodnocení a behaviorální logy z virtuálního vzdělávacího prostředí (VLE). Pro účely této práce byla provedena extrakce a integrace dat ze dvou klíčových entit:

Tabulka studentInfo (Demografický profil):

Tabulka slouží jako statický základ datové sady. Obsahuje unikátní záznamy o 32 593 studentech. Každý záznam definuje profil studenta v době zápisu. Mezi klíčové atributy patří:

- Index sociálního znevýhodnění (IMD): Kategorizace vycházející z adresy bydliště studenta v UK, vyjadřující míru deprivace lokality.
- Nejvyšší dosažené vzdělání: Škála od základního vzdělání až po postgraduální úroveň.
- Studijní zátěž: Definovaná počtem zapsaných kreditů v daném modulu.
- Výsledek (Target): Klasifikace finálního stavu (Pass, Distinction, Fail, Withdrawn).

Tabulka studentVle (Behaviorální stopy):

Soubor obsahuje více než 10 milionů záznamů o interakcích. Každý řádek reprezentuje aktivitu konkrétního studenta v konkrétní den vůči určitému studijnímu materiálu (např. stažení PDF, zhlédnutí videa, odevzdání úkolu). Data jsou zaznamenána jako:

- id_student: Identifikátor studenta.
- id_site: Identifikátor konkrétního materiálu ve VLE.
- date: Relativní den interakce vzhledem k začátku semestru.
- sum_click: Počet kliknutí (interakcí) v daném dni.

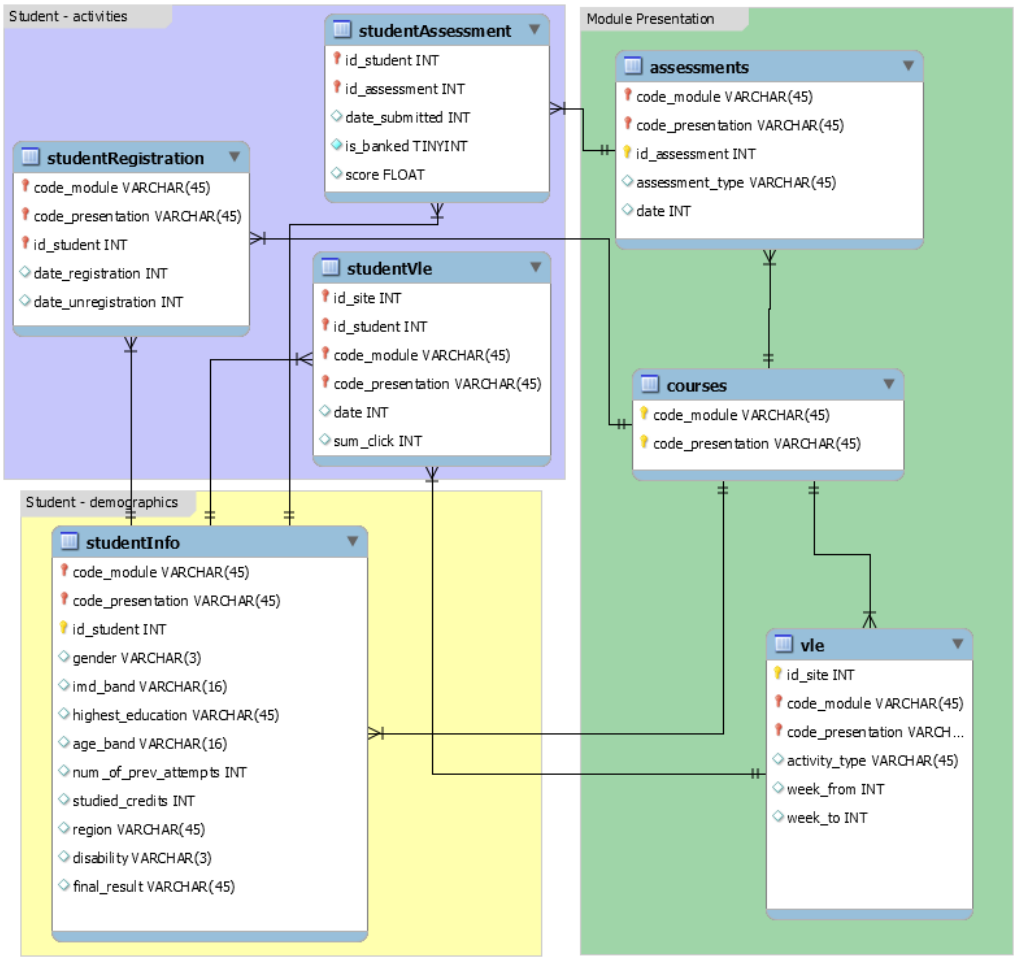
Dataset obsahuje data ze sedmi vybraných modulů (předmětů), které reprezentují široké spektrum akademických disciplín. Moduly lze rozdělit do dvou hlavních domén: společenské vědy (Social Sciences) a technické vědy (STEM). Tabulka 1 shrnuje základní charakteristiku těchto modulů.

Tab. 1: Přehled a kategorizace studijních modulů v datasetu OULAD

code_module (kód modulu)	Obor studia
AAA	Společenské vědy
BBB	Společenské vědy
CCC	STEM (Technické)
DDD	STEM (Technické)
EEE	STEM (Technické)
FFF	STEM (Technické)
GGG	Společenské vědy

Zdroj: vlastní zpracování dle KUZILEK, J., HLOSTA, M. a ZDRAHAL, Z. (2017)

Vzhledem k tomu, že data v studentVle jsou uložena v transakčním formátu (mnoho řádků pro jednoho studenta), bylo nutné provést agregaci na úrovni studenta, aby bylo možné data spojit s tabulkou studentInfo. Sloučení proběhlo pomocí primárního klíče id_student a doplňkového klíče code_module a code_presentation. Tímto procesem vznikla finální analytická tabulka (Analytical Base Table), která pro každého studenta obsahuje unikátní řádek kombinující jeho administrativní předpoklady se sumarizovaným ukazatelem jeho digitální aktivity.



Obr. 2: Schéma propojení tabulek datasetu OULAD

Zdroj: Open University, 2017

Model může kriticky posoudit, zda má na studijní úspěch dominantnější vliv statický profil studenta, tedy jeho demografické pozadí, nebo dynamické aspekty jeho chování reprezentované aktivitou v prostředí VLE. Etický rozměr výzkumu a striktní požadavky na ochranu osobních údajů jsou v rámci celého procesu zajištěny důslednou anonymizací dat, kdy jsou reálná jména a adresy nahrazeny unikátními číselnými identifikátory. Tímto postupem jsou naplněny standardy pro nakládání s citlivými daty v akademickém prostředí.

Výsledkem komplexního procesu čištění, agregace a propojování dat, realizovaného primárně pomocí operací left join na unikátní klíč studenta, je robustní datová sada typu ABT (Analytical Base Table). Tabulka představuje finální a konsolidovaný vstup pro všechny následné statistické operace a modely strojového učení. Celkový rozsah sady činí 32 593 unikátních záznamů, přičemž každý řádek reprezentuje specifickou interakci jednoho studenta v rámci konkrétního studijního modulu.

Unikátnost této sady spočívá především v její multidimenzionalitě, neboť v rámci jediného datového objektu efektivně integruje tři různorodé datové vrstvy. První vrstvu tvoří celkem jedenáct statických prediktorů, které zahrnují demografické faktory, jako jsou pohlaví, věková skupina či region, socioekonomické ukazatele v podobě indexu IMD a nejvyššího dosaženého

vzdělání, a také administrativní data o počtu zapsaných kreditů či předchozích pokusech o studium. Druhou vrstvu představuje klíčový dynamický behaviorální prediktor, tedy agregovaná hodnota sum_click, která v sobě nese komprimovanou a vysoce relevantní informaci o digitální aktivitě studenta v celém průběhu semestru. Poslední vrstvou je pak cílová proměnná vyjádřená binárním klasifikátorem úspěchu či neúspěchu.

Vzniklá analytická sada byla v závěrečné fázi přípravy podrobena důkladné kontrole konzistence. Zjištěná průměrná hodnota aktivity napříč všemi záznamy vykazuje značnou variabilitu a vysokou směrodatnou odchylku. Z pohledu prediktivního modelování se jedná o vysoce žádoucí stav, neboť data v této podobě obsahují dostatečné množství informačního šumu a variability nezbytné pro efektivní diskriminaci mezi úspěšnými a neúspěšnými studenty. Statistický profil ABT tabulky tak vytváří ideální podmínky pro trénování robustních algoritmů strojového učení.

2.3.2 Čištění a úpravy

Kvalita vstupních dat představuje kritický faktor, který přímo podmiňuje úspěšnost a predikční stabilitu jakéhokoliv modelu strojového učení. Fáze předzpracování (Data Preprocessing) v tomto projektu zahrnovala několik komplexních kroků, počínaje agresivní agregací behaviorálních dat. Záznamy o denní aktivitě studentů v systému VLE byly pomocí operace groupby sumarizovány do nové proměnné sum_click. Metrika v sobě koncentruje celkovou míru digitální stopy studenta a slouží jako primární indikátor jeho zapojení do výuky.

```
vle_agg = student_vle.groupby('id_student')['sum_click'].sum().reset_index()
df = student_info.merge(vle_agg, on='id_student', how='left')
df['sum_click'] = df['sum_click'].fillna(0)
df = df.dropna(subset=['final_result'])
df['target'] = df['final_result'].apply(lambda x: 1 if x in ['Pass', 'Distinction'] else 0)
df_raw = df.copy()

print(f"Data zkonsolidována. Počet záznamů: {len(df)}")
```

Obr. 3: Konsolidace dat do proměnné sum_click

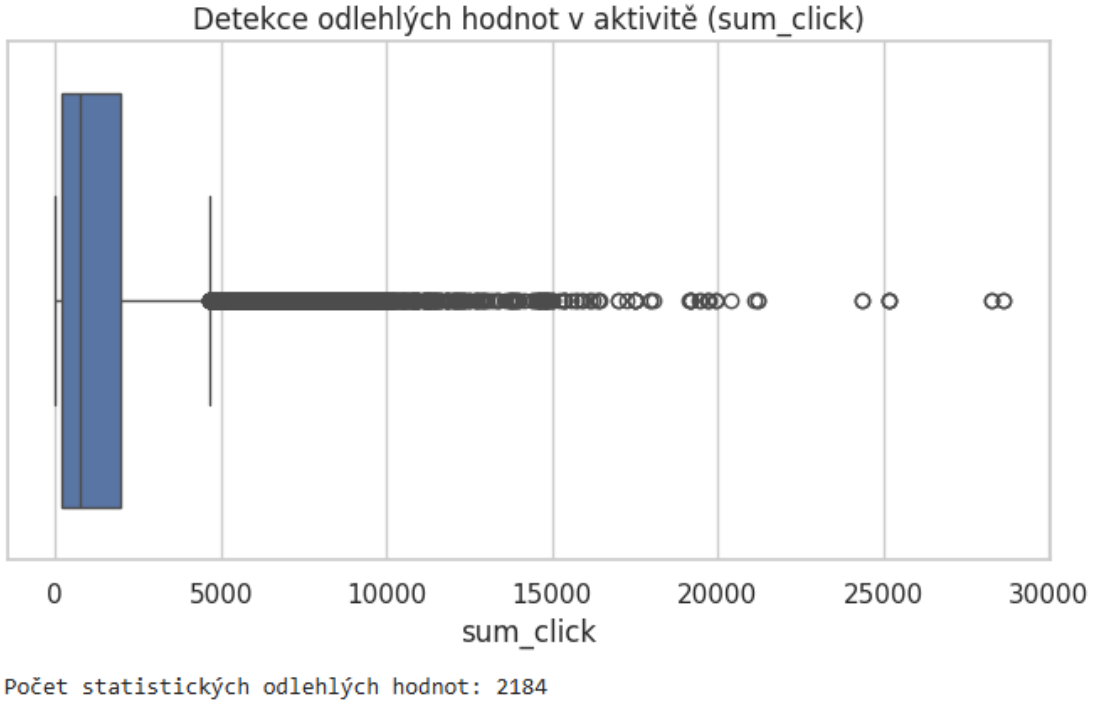
Zdroj: vlastní zpracování

Nedílnou součástí přípravy byla precizní imputace chybějících hodnot. Studenti, kteří se do systému VLE v průběhu semestru vůbec nepřihlásili, generovali v agregované tabulce prázdné záznamy (NaN). Hodnoty byly logicky nahrazeny nulou, což reflektuje reálný stav nulové aktivity. Naopak záznamy, u kterých zcela chyběl definovaný finální výsledek studia, byly z datové sady kompletně odstraněny. Redukce těchto záznamů byla nezbytná, neboť bez validního označení cílové třídy nelze provádět trénování ani validaci predikčního modelu.

Klíčovým rozhodnutím byla transformace původní více třídní proměnné final_result, zahrnující kategorie Pass, Distinction, Fail a Withdrawn, na binární klasifikační problém. Pro účely modelu byly třídy vyjadřující úspěch (Pass, Distinction) zakódovány jako hodnota 1, zatímco kategorie indikující neúspěch nebo předčasné ukončení studia (Fail, Withdrawn) byly sjednoceny pod hodnotu 0. Tímto procesem byl dataset optimalizován pro potřeby binárních klasifikátorů, což zjednodušilo interpretaci výsledků z pohledu včasného varování.

V rámci analytické fáze byla rovněž provedena detekce odlehlých hodnot (outliers) u proměnné sum_click s využitím metody mezikvartilového rozpětí (IQR). Přestože vizualizace na

obrázku 4 potvrdila přítomnost značného množství extrémně aktivních studentů, záznamy nebyly z datové sady odstraněny. Toto rozhodnutí vychází z povahy zvoleného modelu na bázi rozhodovacích stromů (XGBoost), který vykazuje přirozenou rezistenci vůči odlehlým hodnotám. Odstranění těchto dat by navíc vedlo k nežádoucí ztrátě cenných informací o specifických vzorcích chování nejúspěšnějších studentů.



Obr. 4: Detekce odlehlých hodnot v aktivitě (sum_click)

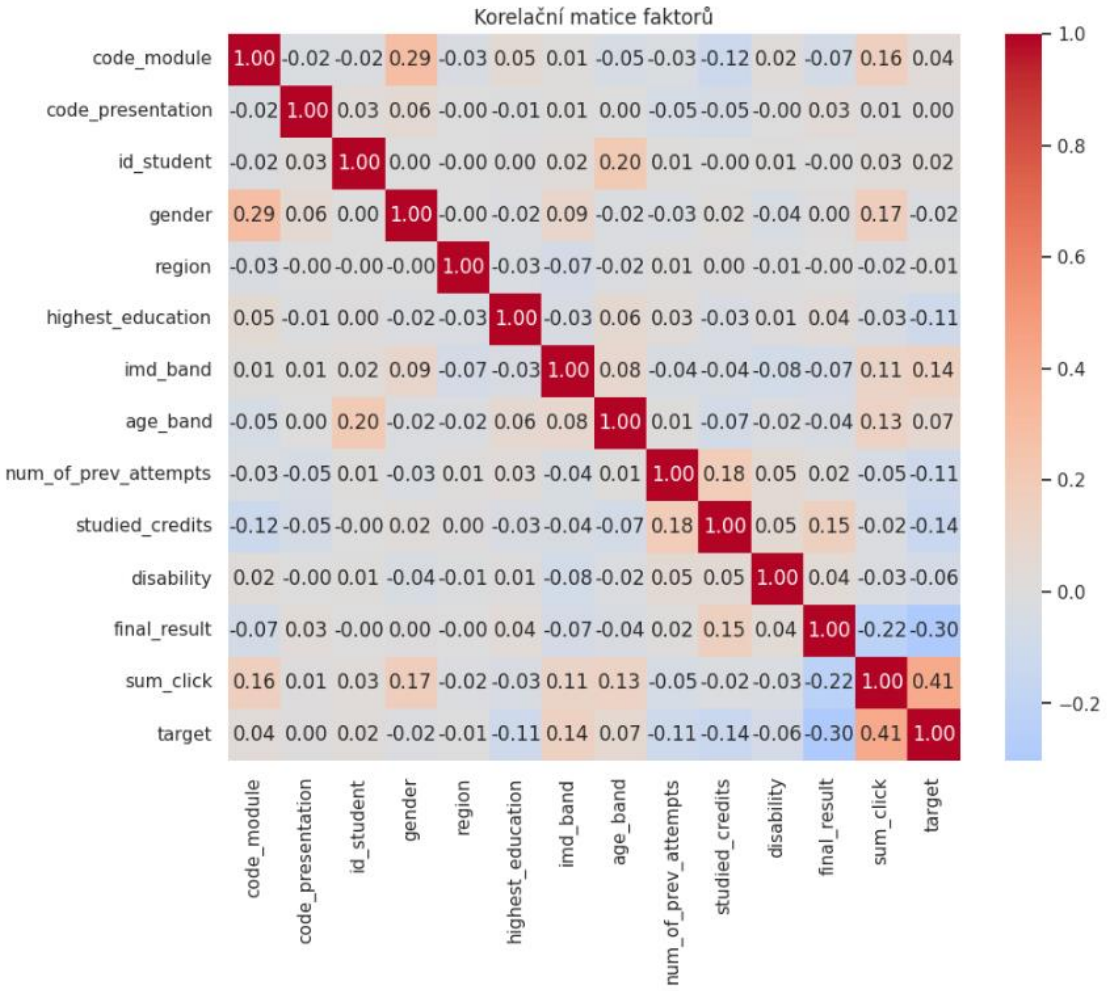
Zdroj: vlastní zpracování

Ústředním prvkem přípravy dat byl proces Feature Engineeringu, konkrétně vytvoření již zmíněné agregované metriky sum_click. V surovém datasetu OULAD je aktivita studenta rozmělněna do tisíců řádků transakčních logů, což je pro klasické modely strojového učení nevhodný formát. Agregace těchto interakcí do jediné proměnné umožňuje modelu zachytit celkovou intenzitu zapojení studenta, označovanou jako student engagement. Kumulativní hodnota byla zvolena jako konceptuální vyjádření reálného studijního úsilí, které se v prostředí online vzdělávání ukazuje jako signifikantně silnější prediktor úspěchu než statické demografické údaje. Ačkoliv moderní analytické metody umožňují i sledování dynamiky aktivity v čase, pro potřeby tohoto robustního modelu byla zvolena právě kumulativní hodnota, která poskytuje stabilní základ pro včasnou a spolehlivou identifikaci rizikových studentů.

2.3.3 Explorativní analýza (EDA)

Cílem explorativní analýzy bylo pochopit vnitřní strukturu dat, identifikovat signifikantní proměnné a ověřit předpoklady pro následné induktivní modelování. Vzhledem k heterogenní povaze datasetu OULAD, který kombinuje demografické údaje s dynamickými záznamy o aktivitě, byla EDA klíčová pro správné nastavení parametrů algoritmů. Základním nástrojem pro identifikaci vztahů byla korelační matice. Jelikož dataset obsahuje řadu kategoriálních proměnných, jako jsou region, dosažené vzdělání nebo výsledná známka, byla aplikována

metoda LabelEncoder. Ta převedla textové hodnoty na numerickou škálu při zachování jejich informační hodnoty, což umožnilo exaktní výpočet korelací mezi všemi atributy.



Obr. 5: Korelační matice a distribuce aktivity

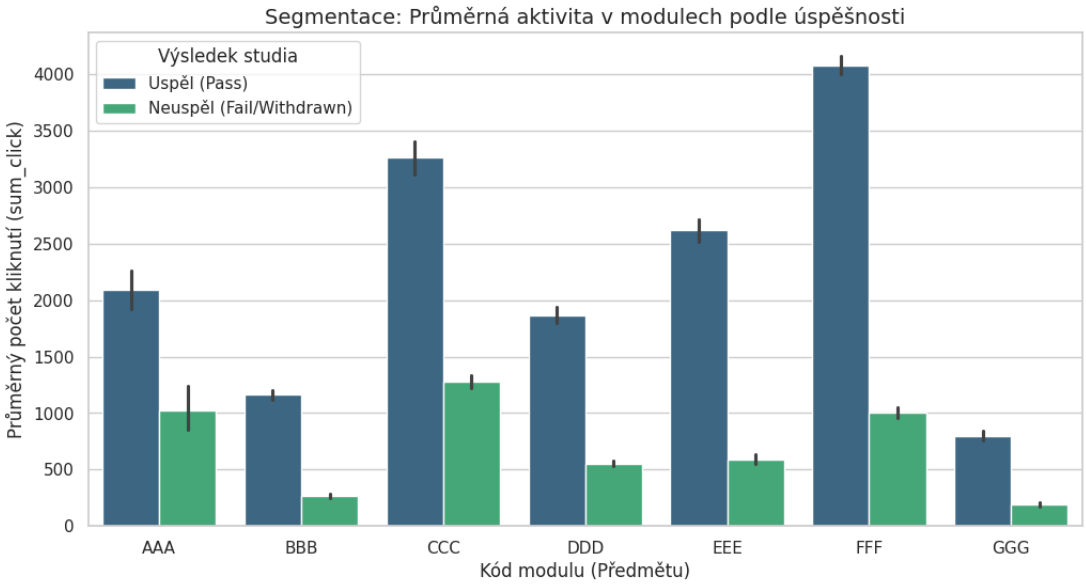
Zdroj: vlastní zpracování

Z vizualizace heatmapy vyplynulo několik zásadních zjištění, z nichž nejvýraznějším je dominance digitální stopy. Proměnná sum_click vykazuje nejsilnější pozitivní korelaci s cílovou proměnnou, což potvrzuje tezi, že kvantita interakcí s virtuálním studijním prostředím (VLE) je primárním prediktorem úspěchu. Dalším zajímavým faktorem je vliv studijní zátěže, kde atribut studied_credits vykazuje mírně negativní korelaci s úspěšností. Jev lze interpretovat jako projev kognitivního přetížení – studenti, kteří si v jednom semestru zapíší nadlimitní množství kurzů, vykazují statisticky vyšší náchylnost k neúspěchu nebo předčasnému ukončení studia. Analýza rovněž ukázala nízkou relevanci statických faktorů, kdy demografické ukazatele jako gender nebo region vykazují vůči výsledku studia korelaci blízkou nule. Pro následné modelování to znamená, že úspěch v online vzdělávání je determinován spíše aktuálním chováním studenta než jeho sociodemografickým pozadím.

Zatímco korelační matice potvrdila globální význam aktivity, detailní segmentace podle jednotlivých studijních modulů odhalila, že kritická hranice aktivity není pro všechny předměty stejná. Vizualizace přináší hlubší vhled do heterogenity studijního prostředí OULAD, zejména v oblasti inter-modulární variability. Z grafu je patrné, že průměrný počet kliknutí u úspěšných

studentů se mezi jednotlivými moduly dramaticky liší. Například v modulu AAA nebo BBB je k úspěchu zapotřebí signifikantně vyšší objem interakcí než v modulech EEE či GGG. Rozdíl mezi moduly reflektuje odlišnou pedagogickou koncepci kurzů, kdy některé jsou orientovány na intenzivní online práci, zatímco jiné kladou větší důraz na samostudium off-line materiálů.

Navzdory těmto rozdílům je ve všech sledovaných modulech patrná konzistence „propasti" v aktivitě. Úspěšní studenti vykazují průměrně dvojnásobnou až trojnásobnou aktivitu oproti studentům neúspěšným, což potvrzuje, že sum_click je univerzálním diskriminátorem bez ohledu na věcné zaměření kurzu. Detailní pohled na segmenty navíc umožňuje identifikaci anomálií, tedy modulů s vysokým rozptylem aktivity u neúspěšných studentů. To naznačuje, že v určitých předmětech nemusí pouhá „přítomnost" v systému garantovat úspěch, pokud není doprovázena kvalitativním zvládnutím látky.



Obr. 6: Segmentace: Průměrná aktivita v modulech podle úspěšnosti

Zdroj: vlastní zpracování

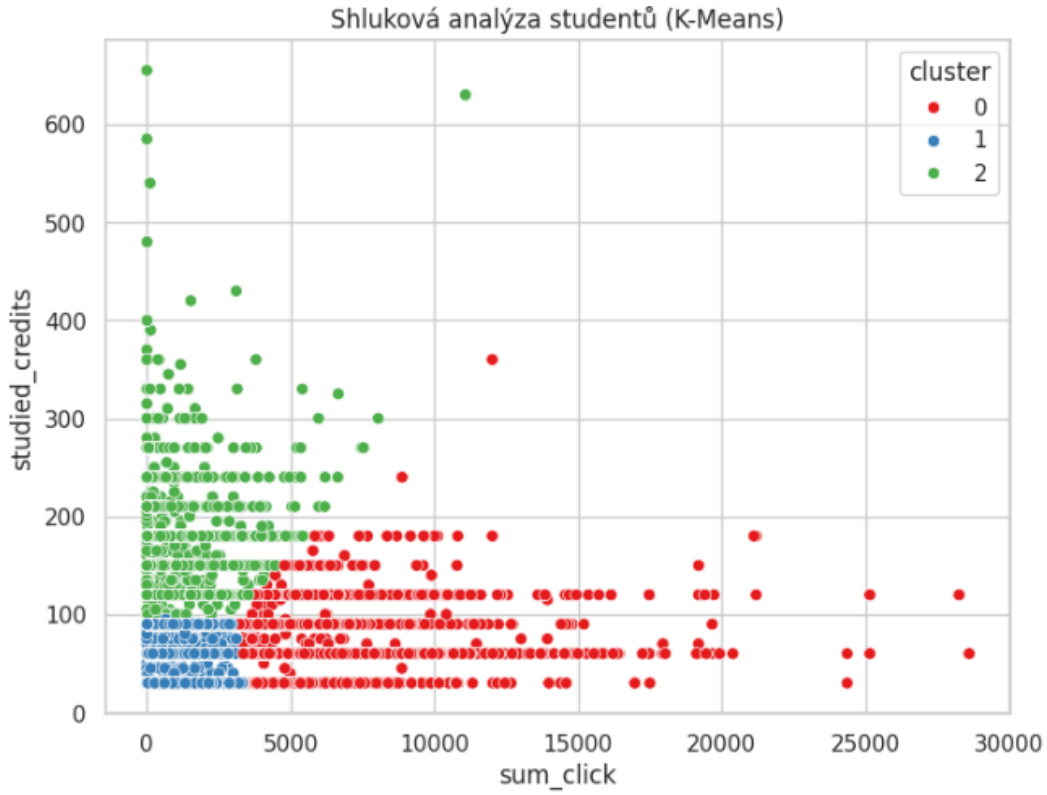
Všechny poznatky měly přímý vliv na fázi induktivního modelování. Potvrdilo se, že pro dosažení vysoké přesnosti predikce, zejména u algoritmu XGBoost, je nezbytné zahrnout code_module jako klíčový kategorický rys. Model se tak „učí", že rizikové chování reprezentované nízkou aktivitou má v každém předmětu jinou absolutní hodnotu. Díky tomu může výsledný dashboard pro koncového uživatele poskytovat relevantnější výsledky v kontextu konkrétního zapsaného předmětu, což zvyšuje praktickou použitelnost celého nástroje.

2.3.4 Učení bez učitele (shlukování)

Před přistoupením k prediktivnímu modelování byla provedena shluková analýza (Clustering) využívající algoritmus K-Means. Účelem bylo odhalit skryté profily studentů na základě dvou klíčových dimenzí: administrativní studijní zátěže (studied_credits) a reálného studijního úsilí (sum_click).

Vzhledem k tomu, že algoritmus K-Means využívá pro výpočet vzdálenosti Eukleidovskou metriku, bylo absolutně nezbytná data nejprve normalizovat pomocí transformace StandardScaler. Bez tohoto kroku by proměnná s řádově vyššími hodnotami (tisíce kliknutí)

algoritmicky zcela potlačila vliv proměnné s nižšími hodnotami (desítky kreditů). Pro zajištění reprodukovatelnosti experimentu byl generátor náhodných čísel zafixován parametrem random_state=42. Algoritmus studenty segmentoval do tří jasně vymezených shluků, které pedagogům umožňují diferencovat přístup (např. identifikovat studenty s vysokou zátěží, ale neadekvátně nízkou aktivitou).



Obr. 7: Vizualizace shluků (K-Means Clustering)

Zdroj: vlastní zpracování

2.3.5 Učení s učitelem (modelování)

Pro hlavní úlohu binární klasifikace byl dataset metodicky rozdělen na trénovací (80 %) a testovací (20 %) sadu s využitím funkce train_test_split. V rámci tohoto procesu byla implementována striktní izolace testovacích dat, což je kritický krok k zamezení problému úniku informací (data leakage). Izolací dat zajišťujeme, že model je evaluován na datech, která během fáze učení nikdy „neviděl“, čímž je zaručena objektivita výsledných metrik a schopnost modelu generalizovat na nové, dosud neznámé studenty.

Do srovnávací analýzy byly zahrnuty tři modely reprezentující odlišné matematické a algoritmické přístupy k řešení klasifikačních úloh. Prvním z nich je Logistická regrese, která v tomto projektu slouží jako základní srovnávací model (baseline). Jedná se o lineární klasifikátor, který predikuje pravděpodobnost příslušnosti k určité třídě pomocí logistické funkce (sigmoid). Ačkoliv vyniká vysokou interpretační schopností, u komplexních behaviorálních dat se předpokládalo, že nedokáže plně zachytit složité nelineární vazby a interakce mezi demografickými faktory a intenzitou aktivity ve VLE.

Druhým analyzovaným modelem je Random Forest (Náhodný les), který reprezentuje rodinu ansámblových metod založených na technice bagging (Bootstrap Aggregating). Algoritmus konstruuje velké množství nezávislých rozhodovacích stromů na náhodných výběrech dat i příznaků. Výsledná predikce je pak tvořena většinovým hlasováním (u klasifikace) všech stromů v lese. Hlasováním efektivně snižuje rozptyl modelu a je velmi odolný vůči přetrénování, což z něj činí robustní nástroj pro analýzu heterogenních dat o studentech.

Třetím a technologicky nejvyspělejším modelem je XGBoost (eXtreme Gradient Boosting). Na rozdíl od náhodného lesa využívá XGBoost techniku gradient boosting, kde jsou rozhodovací stromy stavěny sekvenčně. Každý nový strom se v tomto procesu cíleně zaměřuje na korekci reziduálních chyb, kterých se dopustily stromy předchozí. Implementace navíc využívá pokročilé matematické techniky, jako je regularizace a efektivní výpočet gradientů druhého řádu, což z ní činí jeden z nejvýkonnějších nástrojů pro predikci z tabulkových dat. XGBoost je schopen automaticky identifikovat i velmi jemné nelineární vzorce v chování studentů, které lineární modely obvykle přehlížejí.

Z metodického hlediska bylo nezbytné zajistit konzistenci transformací kategoriálních proměnných. Všechny instance LabelEncoder byly proto uloženy do dedikovaného slovníku (le_dict). Uložení instancí je fundamentální pro následné nasazení modelu do produkčního prostředí v podobě interaktivního dashboardu. Zajišťuje totiž, že libovolný textový vstup od koncového uživatele (např. název regionu nebo kód modulu) bude pro účely predikce zakódován přesně tím číselným identifikátorem, se kterým byl model trénován, čímž je eliminováno riziko sémantických chyb při zpracování dat v reálném čase.

2.3.6 Návrh experimentů a optimalizace hyperparametrů

Experimentální část práce byla koncipována jako strukturovaný proces rozdělený do tří logických etap, které umožnily systematicky sledovat nárůst kvality predikce v závislosti na zvyšující se komplexitě použitých algoritmů. V první fázi, označené jako Experiment 1 – Baseline model, byla jako výchozí bod zvolena logistická regrese. Lineární model sloužil k základnímu ověření hypotézy, zda jsou vztahy v behaviorálních datech studentů jednoduše lineárně oddělitelné. Hlavní výhodou tohoto přístupu je vysoká míra interpretovatelnosti výsledných koeficientů, nicméně jeho zásadní nevýhodou zůstává neschopnost zachytit komplexní nelineární interakce mezi proměnnými, jako je například specifická kombinace věkové skupiny a intenzity aktivity v určitém typu studijního modulu.

Druhá fáze, Experiment 2 – Ensemble modely, byla zaměřena na nasazení pokročilých algoritmů souborového učení, konkrétně metod Random Forest a XGBoost. Volba těchto zástupců nebyla náhodná; zatímco Random Forest využívá principu baggingu (Bootstrap Aggregating) pro snížení rozptylu a minimalizaci rizika přetrénování modelu, XGBoost staví na technice boostingu, která cíleně redukuje zkreslení (bias) tím, že se sekvenčně učí z chyb předchozích iterací. Srovnání těchto dvou přístupů umožnilo identifikovat algoritmus, který nejlépe reflektuje vnitřní strukturu datasetu OULAD.

Závěrečná etapa, Experiment 3 – Optimalizace hyperparametrů, se soustředila na jemné doladění nejúspěšnějšího modelu, kterým se stal XGBoost. Pro nalezení ideální konfigurace byl navržen optimalizační experiment s využitím moderní knihovny Optuna. Namísto tradičního a výpočetně náročného vyčerpávajícího vyhledávání (Grid Search) byl aplikován algoritmus

bayesovské optimalizace, který umožňuje inteligentní a efektivní prohledávání multidimenzionálního prostoru parametrů, jako je maximální hloubka stromů či učící tempo (learning rate), přičemž každá další iterace využívá znalosti získané z pokusů předchozích k dosažení globálního optima.

```
def objective(trial):
    param = {
        'n_estimators': trial.suggest_int('n_estimators', 100, 500),
        'max_depth': trial.suggest_int('max_depth', 3, 9),
        'learning_rate': trial.suggest_float('learning_rate', 0.01, 0.3),
        'subsample': trial.suggest_float('subsample', 0.5, 1.0),
    }
    model = XGBClassifier(**param, eval_metric='logloss')
    model.fit(X_train, y_train)
    return roc_auc_score(y_test, model.predict_proba(X_test)[:, 1])

study = optuna.create_study(direction='maximize')
study.optimize(objective, n_trials=10)

final_model = XGBClassifier(**study.best_params, eval_metric='logloss')
final_model.fit(X_train, y_train)
```

Obr. 8: Implementace objektivní funkce pro Bayesovskou optimalizaci

Zdroj: vlastní zpracování

Metodika testování a evaluace byla postavena na standardizovaném rozdělení dat v poměru 80/20, známém jako Hold-out method. Poměr 80/20 je v oblasti strojového učení považován za optimální, neboť 80 % dat poskytuje dostatečný objem pro učení komplexních vzorců, zatímco zbylých 20 % tvoří zcela nezávislou sadu pro finální a objektivní ověření kvality predikce. Aby se však eliminovalo riziko náhodného zkreslení výsledků vlivem konkrétního náhodného rozdělení, byla během fáze ladění parametrů důsledně využívána technika K-fold cross-validation (křížová validace).

V rámci tohoto procesu byla data rozdělena do pěti bloků (folds), přičemž model byl pětkrát trénován a následně validován, pokaždé na jiné, dříve nepoužité části dat. Výsledná hodnota sledované metriky je pak aritmetickým průměrem těchto pěti běhů, což zaručuje vysokou statistickou stabilitu a robustnost výsledného modelu. Jako hlavní hodnotící kritérium byla zvolena metrika AUC (Area Under the Curve). Na rozdíl od běžné přesnosti (Accuracy), která může být zavádějící v případě nevyváženého zastoupení tříd v datasetu, je AUC imunní vůči nerovnoměrnému podílu úspěšných a neúspěšných studentů. Metrika komplexně hodnotí schopnost modelu správně seřadit studenty podle míry jejich rizika, což je pro praktické nasazení v systému včasného varování klíčový parametr.

2.3.7 Vizualizace uživatelského rozhraní (Tvorba prototypu)

Závěrečnou a klíčovou fází metodiky RAD byla transformace komplexního analytického skriptu do podoby interaktivního webového rozhraní, které je integrováno přímo v prostředí Kaggle. Zvolený přístup umožnil vytvořit funkční most mezi datovou vědou a koncovým uživatelem bez nutnosti opouštět ekosystém Pythonu. Prostřednictvím modulů ipywidgets byl navržen intuitivní

akordeonový layout, který logicky rozděluje funkcionality nástroje do tří hlavních sekcí. První sekce představuje rozhraní pro hromadné zpracování dat, které je navrženo pro efektivní import surových CSV souborů přímo z univerzitních databází. Model v této fázi plně automatizuje proces přípravy: provádí imputaci chybějících hodnot pomocí mediánu, zajišťuje správné kódování kategoriálních proměnných a vyhodnocuje pravděpodobnost úspěchu pro všechny importované záznamy. Výstupem je pak optimalizovaný soubor pro MS Excel s kódováním utf-8-sig, který je připraven k okamžitému použití v administrativě školy.

Druhá část rozhraní se zaměřuje na parametrizaci a dává uživateli unikátní možnost ovlivnit chování predikčního algoritmu „za běhu“. Pomocí interaktivních posuvníků lze definovat klíčové hyperparametry modelu, jako je maximální hloubka stromů nebo jejich celkový počet, a následně iniciovat přetrénování na aktuálních datech. Zmíněná funkce je zásadní pro udržení relevance modelu v čase, kdy se mohou měnit charakteristiky studijních skupin. Celý systém je završen prognostickým dashboardem studenta, což je mřížkové uživatelské rozhraní určené pro okamžitou What-If analýzu.

Dashboard umožňuje studijním poradcům simulovat profil konkrétního studenta a sledovat, jak se mění riziko předčasného ukončení studia při změně vstupních parametrů, jako je počet zapsaných kreditů nebo intenzita aktivity ve VLE. Vizuální indikace míry rizika poskytuje okamžitou zpětnou vazbu, která je nezbytná pro včasnou intervenci. Celé řešení tak demonstruje, jak lze pokročilé metody gradientního boostingu transformovat do srozumitelného a prakticky využitelného nástroje pro podporu akademické úspěšnosti.

```
# Finální sestavení UI
display(widgets.HTML("<br>"))

acc = widgets.Accordion(children=[
    widgets.VBox([upload_widget, output_upload]),
    widgets.VBox([n_est_slider, depth_slider, btn_train, output_train]),
    widgets.VBox([ui_grid, btn_predict, output_pred])
])

acc.set_title(0, 'Nahrávání a integrace externích dat')
acc.set_title(1, 'Experimentální parametrizace a iterace modelu')
acc.set_title(2, 'Interaktivní prediktivní rozhraní pro koncového uživatele')

display(acc)
```

Obr. 9: Tvorba UI rozhraní

Zdroj: vlastní zpracování

3 Výsledky analýz

Cílem této kapitoly je komplexní vyhodnocení natrénovaných modelů a detailní interpretace zjištěných poznatků z pohledu studijní analytiky (Learning Analytics). Fáze představuje kritický bod celého životního cyklu datové vědy, kdy se surové statistické výstupy transformují do podoby srozumitelných a akčních informací, které mohou sloužit jako podklad pro strategické rozhodování vzdělávací instituce. Evaluace se nesoustředí pouze na prosté srovnání predikční přesnosti jednotlivých algoritmů, ale hlouběji analyzuje jejich schopnost identifikovat specifické vzorce chování, které vedou k akademickému neúspěchu.

V rámci vyhodnocení je kladen důraz na robustnost modelů v různých studijních kontextech a na jejich spolehlivost při detekci rizikových studentů. Interpretace výsledků je rozdělena do několika dimenzí, které zahrnují statistickou validaci pomocí pokročilých metrik, analýzu důležitosti jednotlivých příznaků (Feature Importance) a následnou diskusi o limitech a možnostech nasazení těchto systémů v reálném provozu. Kapitola tak přemosťuje matematickou podstatu strojového učení s pedagogickými cíli, kterými jsou především zvýšení retence studentů a včasné poskytnutí studijní podpory těm, kteří vykazují signály možného selhání.

Následující podkapitoly podrobně rozebírají výkonnost baseline modelu logistické regrese v kontrastu s pokročilými ansámblovými metodami. Zvláštní pozornost je věnována modelu XGBoost, který byl podroben hyperparametrické optimalizaci, a jehož schopnost pracovat s nelineárními vztahy v datech o aktivitě ve VLE se ukázala jako klíčová pro dosažení vysoké citlivosti modelu. Výsledky jsou prezentovány formou srovnávacích tabulek a grafických vizualizací, jako jsou ROC křivky a matice záměn, které poskytují ucelený pohled na přesnost a chybovost navrženého řešení.

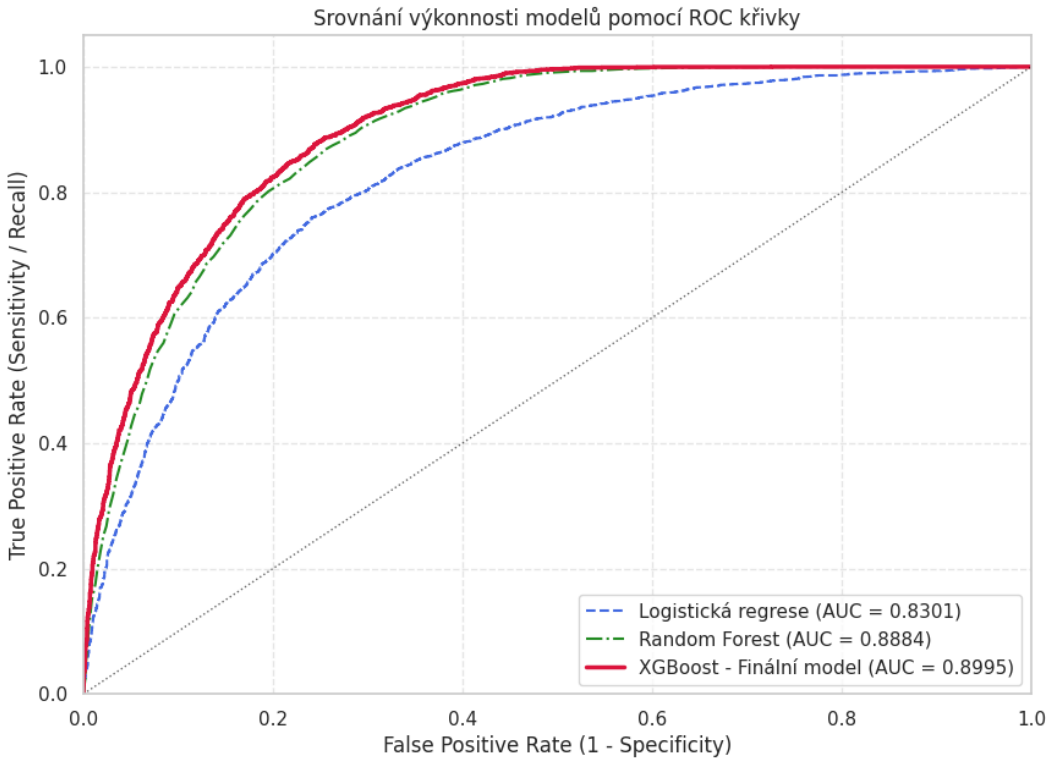
3.1 Srovnání modelů

Výkonnost modelů byla evaluována na izolované testovací sadě pomocí metriky AUC (Area Under the ROC Curve). Metrika AUC byla upřednostněna před pouhou přesností (Accuracy), neboť poskytuje robustnější hodnocení výkonu modelu bez ohledu na případnou nevyváženost tříd.

Výsledky srovnávacího experimentu:

- Logistická regrese (Lineární model): 0.8301
- Random Forest (Ensemble Bagging): 0.8884
- XGBoost (Optimalizovaný Ensemble Boosting): 0.8995

Výsledky potvrzují počáteční předpoklad. Zatímco logistická regrese poskytuje solidní, avšak limitovaný výkon, optimalizovaný XGBoost se přiblížil hranici 0.90 AUC. Z vědeckého hlediska je hodnota považována za vynikající schopnost modelu diskriminovat mezi třídami.

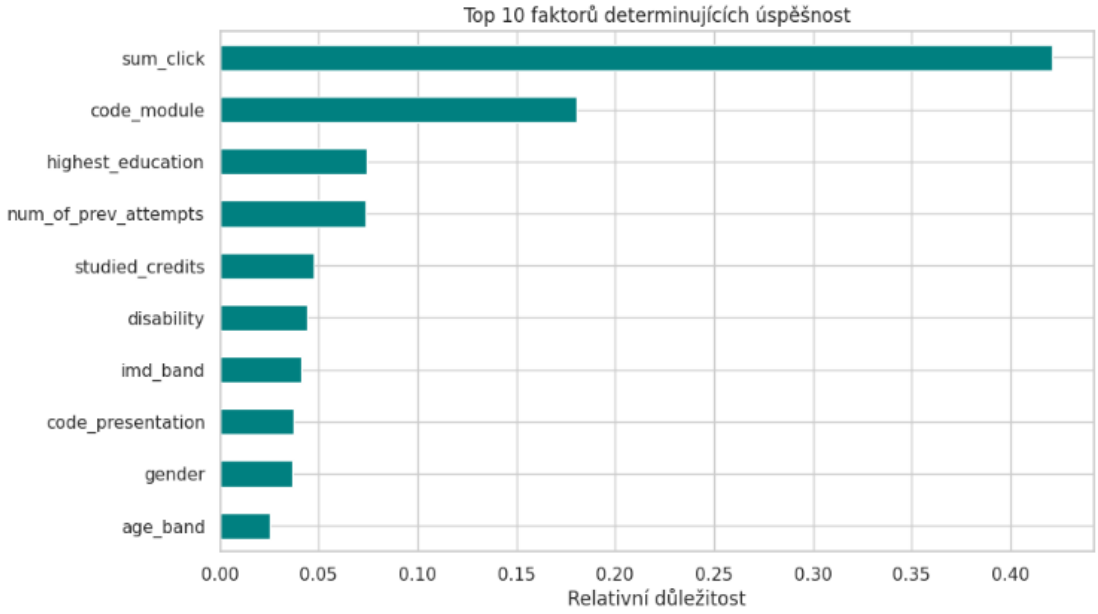


Obr. 10: ROC křivka (Receiver Operating Characteristic)

Zdroj: vlastní zpracování

3.2 Interpretace modelu (Feature Importance)

Moderní datová analytika ve vzdělávání vyžaduje, aby predikční modely nebyly pouze vysoce přesné (tzv. black-box), ale také transparentní a interpretovatelné (tzv. white-box). Pro hlubší pochopení rozhodovacích procesů modelu XGBoost byly extrahovány váhy důležitosti jednotlivých faktorů (feature_importances_). Hodnoty určují, jakým dílem přispěla konkrétní proměnná k redukci chyby predikce a ke zvýšení čistoty rozhodovacích uzlů napříč všemi vytvořenými stromy.



Obr. 11: Důležitost příznaků u modelu XGBoost

Zdroj: vlastní zpracování

Graf relativní důležitosti deseti nejvýznamnějších faktorů odhalil naprosto dominantní postavení proměnné sum_click. Výsledek přináší zásadní akademický i praktický závěr: ačkoliv demografické faktory jako věk, region nebo úroveň sociálního znevýhodnění (index IMD) hrají v životě studenta určitou roli, pro jeho konečný úspěch v online vzdělávacím prostředí je zcela určující jeho reálná, průběžná aktivita ve virtuálním studijním systému (VLE). Z pohledu univerzity to znamená, že efektivní systém včasného varování (Early Warning System) musí být primárně napojen na živé protokoly o interakcích studentů s digitálními materiály, nikoliv pouze na jejich statický administrativní profil vyplněný při podání přihlášky.

Druhým nejvýznamnějším faktorem se s odstupem stala proměnná code_module, tedy kód konkrétního studijního předmětu. Poznatek potvrzuje zjištění z explorativní analýzy o vysoké inter-modulární variabilitě. Model XGBoost správně identifikoval, že „rizikové chování“ nelze definovat univerzálně pro celou univerzitu, ale je nutné jej posuzovat v kontextu konkrétního kurzu. Například nízká aktivita v technicky orientovaném modulu (STEM) může predikovat selhání mnohem dříve a s vyšší jistotou než identická úroveň aktivity v modulu společenskovědním, kde je těžiště práce více přesunuto k samostudiu offline textů.

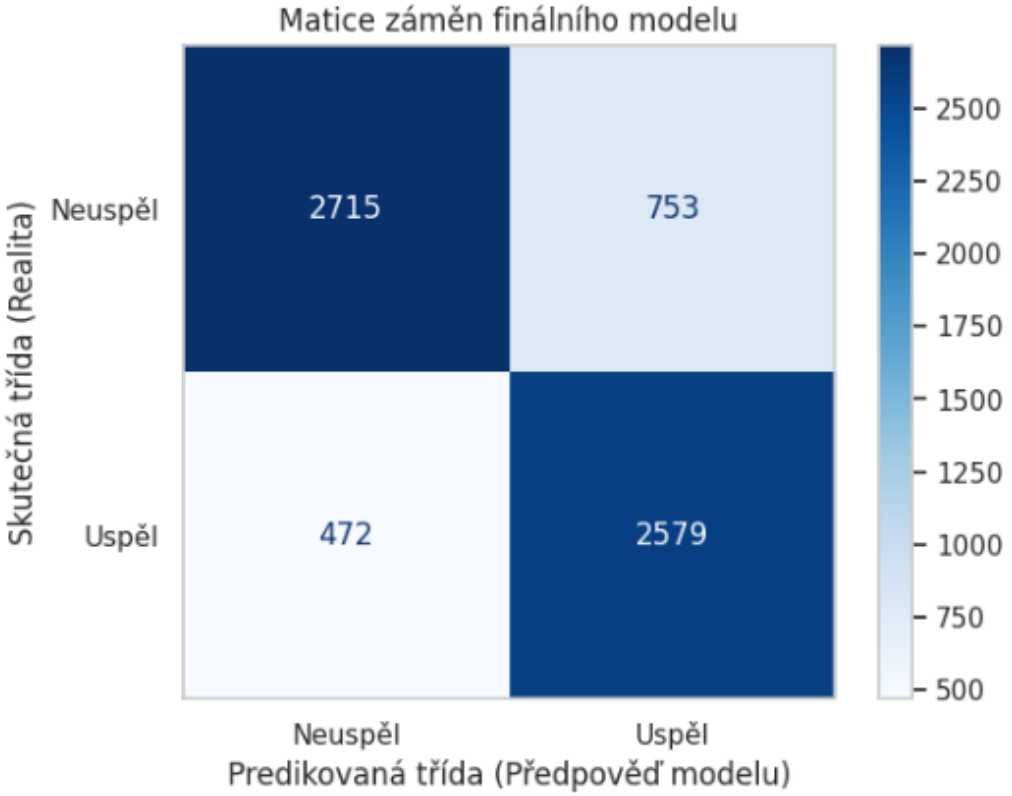
Třetím kritickým prediktorem v pořadí bylo nejvyšší dosažené vzdělání studenta (highest_education). Proměnná v modelu funguje jako proxy ukazatel pro studijní návyky a připravenost na vysokoškolský typ studia. Ukazuje se, že předchozí akademická zkušenost významně ovlivňuje schopnost studenta adaptovat se na specifika online výuky a efektivně si rozvrhnout studijní úsilí. Kombinace těchto tří faktorů – intenzity aktivity, kontextu předmětu a předchozího vzdělání – tvoří v modelu XGBoost silný predikční trojúhelník, který umožňuje s vysokou přesností oddělit úspěšné trajektorie od těch rizikových.

Hierarchie faktorů naznačuje, že i když má instituce omezené možnosti, jak ovlivnit vstupní vzdělání studentů, má obrovský prostor pro intervenci v oblasti jejich digitální aktivity. Pokud systém včasného varování identifikuje pokles v sum_click v rámci konkrétního code_module,

může tutor zasáhnout dříve, než se studijní deficit stane nezvratným. Právě schopnost modelu extrahovat akční znalost z hromady transakčních dat je hlavním přínosem předkládané práce.

3.3 Matice záměn (Confusion Matrix) a limity aplikace

Pro zhodnocení praktického nasazení modelu do reálného pedagogického procesu byla vygenerována matice záměn (Confusion Matrix), která detailně rozkládá distribuci správných predikcí a specifických typů chyb. Vizualizace je nezbytná pro pochopení, jak se model chová v hraničních situacích a jakou důvěru do něj může studijní poradce vložit. Při standardní klasifikační prahu (threshold) 0,5 finální model rozdělil testovací sadu do čtyř základních kvadrantů, které definují úspěšnost predikčního algoritmu.



Obr. 12: Matice záměn finálního modelu

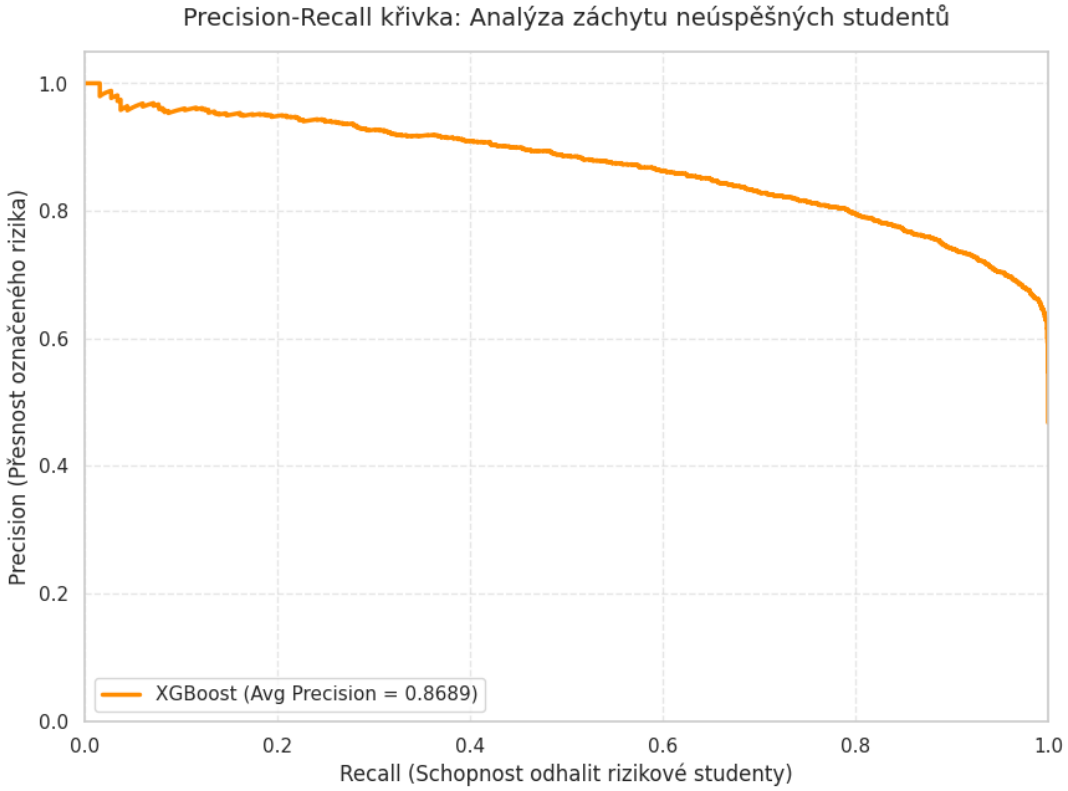
Zdroj: vlastní zpracování

První skupinu tvoří správně predikovaný úspěch (True Positives), tedy studenti, kteří studium řádně dokončili a model je v souladu s realitou identifikoval. Z pohledu strategického řízení školy je však nejdůležitější skupinou správně predikované riziko (True Negatives). Jde o studenty, kteří směřovali k neúspěchu a model jejich ohrožení odhalil včas na základě jejich nízké aktivity. Právě u této skupiny se otevírá prostor pro proaktivní pedagogickou intervenci, kdy mohou tutoři nabídnout cílenou podporu dříve, než student kurz předčasně ukončí.

Analýza chybových stavů odhalila dva typy selhání s odlišným dopadem na akademickou praxi. Falešně pozitivní chyby (chyba I. typu) nastávají v momentě, kdy model studenta označí štítkem „Uspěl“, ačkoliv student ve skutečnosti studium nedokončil. Z hlediska nasazení systému

včasného varování je chyba nejvíce nebezpečná, neboť pokud systém selže v detekci ohroženého studenta, ten neobdrží potřebnou konzultaci a škola ho s vysokou pravděpodobností ztratí. Naopak falešně negativní chyby (chyba II. typu), kdy model predikuje neúspěch u studenta, který nakonec prospěje, jsou v akademickém prostředí akceptovatelnější. Znamenají pouze to, že tutor věnuje čas i studentovi, který by látku zvládl sám, což sice představuje mírnou ztrátu efektivity, ale nemá fatální dopad na celkovou retenci.

Pro hloubkovou analýzu výkonu byly vypočteny metriky Precision (přesnost) a Recall (úplnost/citlivost), které mají v kontextu systémů včasného varování specifický význam. Citlivost udává, jaké procento skutečně neúspěšných studentů dokázal model v datové sadě odhalit, zatímco přesnost určuje, kolik z označených „rizikových" studentů by v realitě skutečně neuspělo. Vzájemný vztah těchto metrik ilustruje Precision-Recall křivka, která potvrzuje stabilitu modelu napříč různými rozhodovacími prahy.



Obr. 13: Precision-Recall Křivka

Zdroj: vlastní zpracování

Z pedagogického hlediska je klíčová právě analýza kritických selhání. Situace, kdy model studenta chybně označí za úspěšného (falešně negativní výsledek z pohledu detekce rizika), znamená, že škola nezasáhne a o studenta přijde. Opačný případ, kdy je student označen za rizikového nadbytečně, představuje pouze neefektivní vynaložení zdrojů, například formou automatizovaného e-mailu, který však studentovi v učebním procesu nijak neuškodí. Finální model XGBoost byl proto laděn s prioritou na maximalizaci citlivosti při zachování rozumné míry přesnosti. V rámci dalšího rozvoje tohoto softwarového prototypu lze uvažovat o umělém posunutí rozhodovací hranice tak, aby se minimalizovala především nebezpečná chyba I. typu, i za cenu mírného zvýšení počtu falešných poplachů.

3.4 Omezení práce a diskuse

Při interpretaci dosažených výsledků a hodnocení úspěšnosti predikčního modelu je nezbytné vzít v úvahu několik limitujících faktorů, které definují hranice použitelnosti navrženého řešení. Prvním zásadním omezením jsou regionální specifika použitého datasetu. Jelikož soubor OULAD pochází z britského univerzitního prostředí, existují signifikantní rozdíly v sociokulturním zázemí studentů a v celkovém systému hodnocení oproti českému vysokému školství. Skutečnost může částečně ovlivnit přenositelnost modelu na lokální podmínky, kde se mohou lišit vzorce chování studentů i administrativní procesy spojené se zápisem do studia.

Dalším faktorem je samotná povaha sledované aktivity. Metrika sum_click poskytuje exaktní kvantitativní vyjádření digitální stopy, tedy kolikrát student s daným materiálem interagoval, nicméně nedokáže zachytit kvalitativní rozměr tohoto procesu. Model v současné podobě nedokáže rozlišit mezi hlubokým a soustředěným studiem studijních opor a pouhým mechanickým proklikáváním stránek v systému VLE bez reálného kognitivního zapojení. V neposlední řadě je nutné zmínit časový aspekt analýzy. Predikce v této práci využívá agregovaná data za celé studijní období, což poskytuje robustní základ pro pochopení faktorů úspěchu, ale pro reálné nasazení v režimu včasného varování (Early Warning System) by bylo žádoucí model v budoucnu modifikovat tak, aby pracoval s inkrementálními daty dostupnými například již v prvních týdnech semestru.

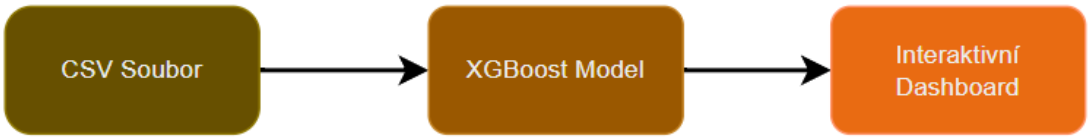
Navzdory těmto omezením dosažené výsledky jasně naznačují, že vzdělávací instituce by měly aktivně monitorovat digitální stopu studentů od samého počátku kurzu. Vytvořený interaktivní prototyp demonstruje, jak může moderní datová věda sloužit jako základ pro dashboardy studijních poradců. Ti mohou díky predikované pravděpodobnosti úspěchu efektivně prioritizovat pomoc studentům, u nichž hodnota klesne pod kritickou hranici 40 %. Takto nastavený systém umožňuje adresně směřovat omezené personální kapacity poradců tam, kde je riziko studijní neúspěšnosti nejvyšší, a tím proaktivně předcházet předčasnému ukončení studia.

Diskuse výsledků v kontextu praxe dále potvrzuje, že i přes naprostou dominanci aktivity nelze v predikčních schématech zcela ignorovat administrativní faktory. Jako nejsilnější indikátor včasného odchodu ze studia (drop-out) se jeví specifická kombinace vysokého počtu zapsaných kreditů a nízké počáteční aktivity ve virtuálním studijním prostředí. Poznatek představuje cenný podklad pro metodiku studijních poradců, kteří by se měli v rámci preventivních opatření zaměřit na studenty s vysokou studijní zátěží hned v úvodních čtrnácti dnech semestru. Integrace těchto analytických zjištění do poradenských procesů na univerzitách by mohla významně přispět ke zvýšení studijní retence a celkové spokojenosti studentů.

4 Závěrečné hodnocení

4.1 Shrnutí hlavních přínosů vzhledem ke zkoumanému problému

Předkládaná práce se zabývala identifikací faktorů ovlivňujících studijní úspěšnost v prostředí distančního vzdělávání na základě analýzy datasetu OULAD. Hlavním přínosem práce je potvrzení hypotézy, že digitální aktivita studenta ve virtuálním vzdělávacím prostředí (VLE) je statisticky významnějším prediktorem úspěchu než jeho demografické či socioekonomické zázemí.

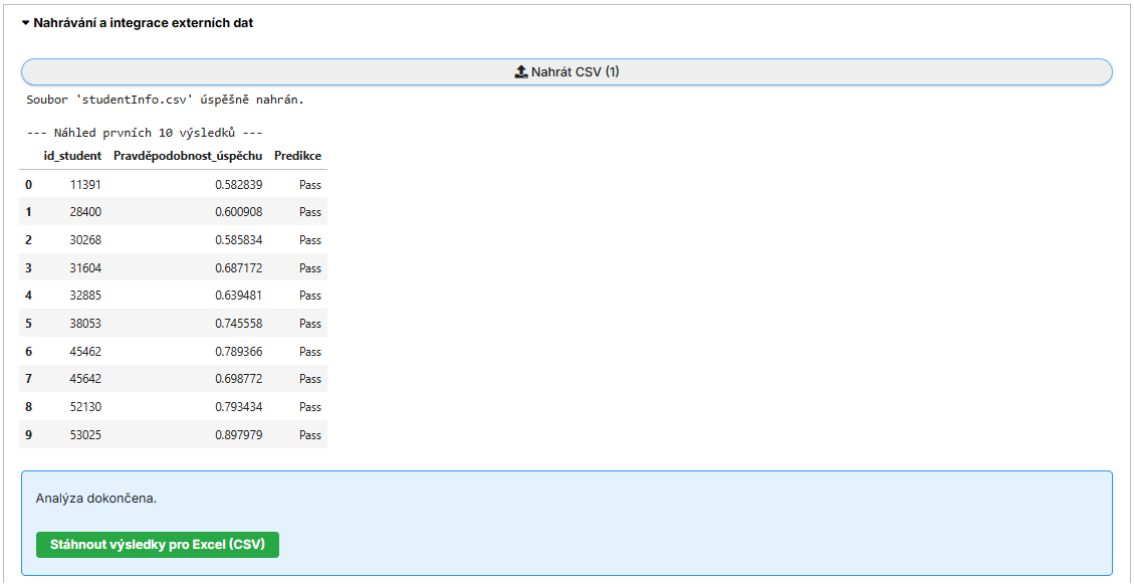


Obr. 14: Workflow diagram aplikace

Zdroj: vlastní zpracování

Z technického hlediska práce přinesla optimalizovaný klasifikační model založený na algoritmu XGBoost, který dosáhl vysoké přesnosti (AUC ≈ 0.90). Praktickým výstupem je funkční softwarový prototyp vyvinutý metodikou RAD. Nástroj umožňuje nejen hromadnou predikci rizikových studentů z databázových exportů, ale nabízí i interaktivní rozhraní pro pedagogy, které slouží jako systém včasného varování (Early Warning System).

Praktickým vyústěním práce je modulární rozhraní (viz Obrázky 15 až 17), které provádí uživatele celým procesem od importu surových dat přes ladění parametrů algoritmu až po finální interpretaci individuálního rizika studenta.



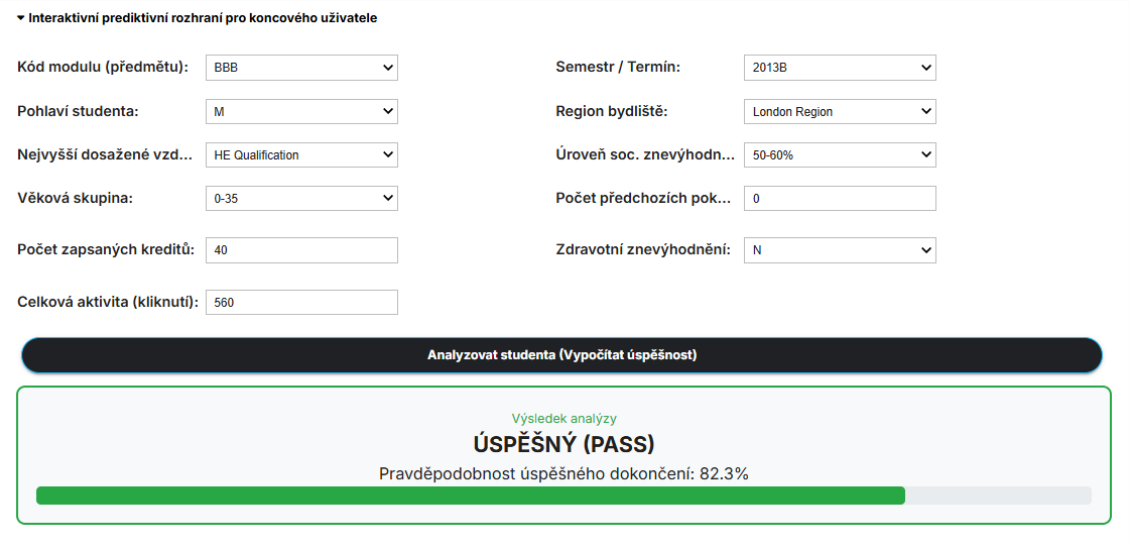
	id_student	Pravděpodobnost_úspěchu	Predikce
0	11391	0.582839	Pass
1	28400	0.600908	Pass
2	30268	0.585834	Pass
3	31604	0.687172	Pass
4	32885	0.639481	Pass
5	38053	0.745558	Pass
6	45462	0.789366	Pass
7	45642	0.698772	Pass
8	52130	0.793434	Pass
9	53025	0.897979	Pass

Obr. 15: Uživatelské rozhraní pro hromadný import externích dat

Zdroj: vlastní zpracování

Obr. 16: Rozhraní pro experimentální parametrizaci a interaktivní ladění modelu XGBoost

Zdroj: vlastní zpracování



Obr. 17: Prediktivní dashboard pro individuální analýzu studijního rizika

Zdroj: vlastní zpracování

4.2 Míra naplnění cílů a výzkumná omezení

V rámci projektu byl úspěšně realizován kompletní proces dolování znalostí z dat v souladu s metodikou CRISP-DM, což zahrnovalo všechny klíčové fáze od prvotní explorativní analýzy a shlukování dat až po finální induktivní modelování a jeho následnou integraci do funkčního softwarového prototypu. Navzdory vysoké úspěšnosti a predikční stabilitě výsledného modelu je však nezbytné kriticky reflektovat určitá výzkumná omezení, která definují rámec interpretace dosažených výsledků.

Prvním z těchto omezení je geografická a kulturní specifičnost použitého zdroje. Data pocházejí z britského prostředí Open University, což znamená, že přenositelnost modelu na české vysoké školství může být ovlivněna odlišnou strukturou semestru, jinou dynamikou studijních skupin i odlišným způsobem průběžného a závěrečného hodnocení. Dalším významným faktorem je diskrepance mezi kvantitou a kvalitou sledované aktivity. Hlavní prediktor sum_click měří primárně objem digitálních interakcí, avšak model sám o sobě nedokáže rozlišit mezi hlubokým, kognitivně náročným studiem materiálů a pouhým povrchním proklikáváním stránek. Limit může v praxi vést k falešně pozitivním predikcím u studentů, kteří vykazují vysokou, ale ve výsledku neefektivní aktivitu v systému VLE.

V neposlední řadě je nutné zmínit absenci predikce v reálném čase jako jeden z limitů současné verze analýzy. Práce operovala s agregovanými daty za celé studijní období, což poskytuje cenný retrospektivní vhled do faktorů úspěšnosti, ale pro dosažení maximální efektivity v reálném provozu by bylo zapotřebí model dynamicky adaptovat. Budoucí rozvoj by se měl zaměřit na tvorbu predikcí v konkrétních časových milnících semestru, například již ve čtvrtém týdnu výuky na základě dosavadního vývoje aktivity. Taková úroveň granularity by umožnila ještě přesnější cílení intervenčních opatření a posílila by roli datové vědy jako nástroje pro aktivní podporu studentů v průběhu celého jejich akademického cyklu.

4.3 Doporučení pro další výzkum

Na základě dosažených výsledků a identifikovaných omezení se pro budoucí výzkum v oblasti Learning Analytics nabízí několik perspektivních směrů, které mohou dále posunout efektivitu predikčních systémů. Prvním z nich je hlubší analýza časových řad, která by nahradila dosavadní práci s kumulativní aktivitou zkoumáním dynamických trendů v čase. Identifikace specifických vzorců, jako je náhlý pokles aktivity v určité fázi semestru, může sloužit jako mnohem citlivější indikátor blížícího se předčasného ukončení studia než prostý součet interakcí za celé období, což by umožnilo modelu lépe rozlišit mezi studenty, kteří svou aktivitu pouze efektivněji rozkládají, a těmi, kteří reálně ztrácejí o studium zájem.

Další významné zpřesnění predikcí by mohla přinést integrace metod zpracování přirozeného jazyka (NLP). Jelikož dataset OULAD neobsahuje přímou textovou komunikaci, budoucí výzkum by se měl zaměřit na analýzu sentimentu a obsahu diskusních fór či e-mailové korespondence. Taková analýza by mohla výrazně zvýšit úspěšnost modelu u studentů, kteří vykazují nižší digitální stopu v rámci studijních materiálů, ale o to intenzivněji a s konkrétními dotazy komunikují s vyučujícími či spolužáky. Propojení behaviorálních dat s lingvistickou analýzou by tak vytvořilo mnohem komplexnější obraz o studijním procesu jednotlivce.

V oblasti důvěryhodnosti a praktické použitelnosti modelů by měl další vývoj směřovat k metodám vysvětlitelné umělé inteligence (XAI), jako jsou techniky SHAP či LIME. Cílem je transformovat systém z podoby „černé skříňky“ do nástroje, který vyučujícím neposkytuje pouze anonymní číslo pravděpodobnosti úspěchu, ale generuje konkrétní textová vysvětlení rizikových faktorů. Učitel by tak mohl obdržet informaci, že daný student je v riziku například z důvodu nulové aktivity v posledních deseti dnech v kombinaci s vysokou studijní zátěží, což by vedlo k mnohem adresnější a efektivnější podpoře.

Konečným cílem budoucího snažení by měla být plná implementace navrženého prototypu do systémů pro řízení výuky (LMS), jako je například platforma Moodle. Transformace analytického nástroje do podoby nativního zásuvného modulu (pluginu) by umožnila automatizované generování upozornění pro studijní poradce v reálném čase. Takové propojení datové vědy s každodenní pedagogickou praxí by vytvořilo ucelený ekosystém pro podporu studijní úspěšnosti, kde by predikční modely nesloužily pouze k retrospektivní analýze, ale staly by se aktivním prvkem prevence studijního neúspěchu v moderním vysokoškolském prostředí.

Závěr

Cílem bakalářské práce byla aplikace úplného životního cyklu datové vědy na problematiku analýzy studijní úspěšnosti s využitím datasetu OULAD. V teoretické části byly ukotveny principy metodiky CRISP-DM a specifika oboru Learning Analytics. Praktickým přínosem práce je návrh a realizace predikčního modelu založeného na algoritmu XGBoost, jehož hyperparametry byly precizně laděny pomocí Bayesovské optimalizace v knihovně Optuna.

Provedená analýza dat potvrdila, že klíčovým faktorem predikce úspěchu není pouze demografické zázemí studenta, ale především jeho průběžná interaktivita ve virtuálním studijním prostředí (VLE), reprezentovaná zejména agregovanou metrikou sum_click. Model prokázal vysokou schopnost identifikovat rizikové studenty již v raných fázích semestru, což je kritické pro včasnou pedagogickou intervenci. Díky využití agilní metodiky Rapid Application Development (RAD) se podařilo technické výstupy modelu transformovat do podoby interaktivního funkčního prototypu dashboardu, který činí výsledky analýzy srozumitelnými i pro uživatele bez hluboké znalosti datové vědy.

Důležitou součástí práce bylo hodnocení modelu pomocí matice záměn a metriky AUC, což potvrdilo robustnost zvoleného algoritmu. Interpretace výsledků ukázala, že kombinace pokročilého strojového učení a doménových znalostí z oblasti vzdělávání umožňuje efektivně identifikovat vzorce chování vedoucí k neúspěchu.

Závěrem lze konstatovat, že datová věda představuje silný nástroj pro zvyšování retence studentů a kvality vysokoškolského vzdělávání. Budoucí vývoj v této oblasti by měl směřovat k integraci těchto modelů přímo do prostředí LMS (např. Moodle) v reálném čase a k širšímu využití metod vysvětlitelné umělé inteligence (XAI), které by pedagogům poskytly ještě hlubší vhled do individuálních potřeb jednotlivých studentů.

Seznam použité literatury

AKIBA, Takuya et al. Optuna: A Next-generation Hyperparameter Optimization Framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discoveryand Data Mining*. 2019. Dostupné z: https://doi.org/10.1145/3292500.3330701

CIELEN, M., A. MEYSMAN a M. ALI. *Introducing Data Science: Big Data, Machine Learning, and More, using Python Tools*. Shelter Island: Manning Publications, 2016. ISBN 978-1-63343-003-7.

DHAR, V. Data Science and Prediction. *Communications of the ACM* [online]. 2013, roč. 56, č. 12, s. 64–73 [cit. 2025-12-05]. ISSN 0001-0782. Dostupné z: https://doi.org/10.1145/2500499

FAWCETT, T. a F. PROVOST. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. Sebastopol: O'Reilly Media, 2013. ISBN 978-1-4493-6132-7.

GOOGLE. *Gemini* [velký jazykový model]. Verze 1.5 Pro. 14. 5. 2024. Dostupné z: https://gemini.google.com. [cit. 2026-01-08].

CHAPMAN, P. et al. *CRISP-DM 1.0: Step-by-step Data Mining Guide*. Chicago: SPSS, 2000.

CHEN, Tianqi a Carlos GUESTRIN. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, s. 785–794. Dostupné z: https://doi.org/10.1145/2939672.2939785

KAGGLE. *Kaggle: Your Home for Data Science* [online]. 2024 [cit. 2026-01-08]. Dostupné z: https://www.kaggle.com

KIMBALL, R. a M. ROSS. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. 3. vyd. Hoboken: Wiley, 2013. ISBN 978-1-118-53080-1.

KUMAR, M. a V. VIJAYAN. Educational Data Mining and Learning Analytics: Summary, Challenges and Future Prospects. *International Journal of Education and Development using ICT* [online]. 2020, roč. 16, č. 3, s. 121–135 [cit. 2025-12-06]. ISSN 1814-0556. Dostupné z: https://eric.ed.gov/?id=EJ1275464

KUZILEK, J., M. HLOSTA a Z. ZDRAHAL. Open University Learning Analytics dataset. *Scientific Data* [online]. 2017, roč. 4, č. 170171 [cit. 2026-01-08]. ISSN 2052-4463. Dostupné z: https://doi.org/10.1038/sdata.2017.171

MACFADYEN, L. P. a S. DAWSON. Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education* [online]. 2010, roč. 54, č. 2, s. 588–599 [cit. 2025-12-10]. ISSN 0360-1315. Dostupné z: https://doi.org/10.1016/j.compedu.2009.09.008

MARTIN, James. *Rapid Application Development*. New York: Macmillan Publishing Co., 1991. ISBN 0-02-376775-8.

OPEN UNIVERSITY. *Open University Learning Analytics Dataset (OULAD)* [online]. 2017 [cit. 2026-01-08]. Dostupné z: https://analyse.kmi.open.ac.uk/open-dataset

PEDREGOSA, Fabian et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011, roč. 12, s. 2825–2830.

PRESS, G. A Very Short History of Data Science. *Forbes* [online]. 28. 5. 2013 [cit. 2025-12-01]. Dostupné z: https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/

SIEMENS, G. a P. LONG. Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review* [online]. 2011, roč. 46, č. 5, s. 30–40 [cit. 2025-12-11]. Dostupné z: https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education

Přílohy

Kaggle notebook dostupný na adrese: https://www.kaggle.com/code/pavelcerny/anal-za-studijn-sp-nosti-v-r-mci-oulad-datasetu