Príloha A

Plán a zhodnotenie práce na projekte

V prílohe nájdete detailný plán práce, ktorý je rozčlenený na dve hlavné fázy - zimný a letný semester. Tento dokument obsahuje prehľad aktivít a cieľov stanovených pre obidve obdobia bakalárskeho projektu, ako aj hodnotenie a reflexiu plnenia týchto plánov v priebehu oboch semestrov.

A.1. Plán práce na zimný semester

Plán práce na zimný semester sme si rozdelili na 5 častí.

Týždeň	**Plán práce pre dané obdobie**
1 - 3	Štúdium a hľadanie hodnotných zdrojov, ktoré sa venujú stylometrií
4 - 5	Analýza strojového učenia a rôzdnych algoritmov
6 - 8	Analýza existujúcich riešení profilácie autorstva so zameraním na vek a pohlavie
9 - 10	Záver analýzy a tvorba hypotézy
11 - 12	Zapracovanie nedostatkov podľa pokynov vedúceho

Tabuľka A.1: Plán práce na zimný semester

A.2. Zhodnotenie plánu práce na zimný semester

Na úvode semestra sme sa intenzívne venovali rozsiahlemu štúdiu literatúry a odborných zdrojov zameraných na stylometriu. Tento komplexný prieskum bol kľúčový pre získanie detailného a hlbokého pochopenia témy, ktorá je komplexná a pútavá. Vďaka týmto rozsiahlym predchádzajúcim štúdiám sme nadobudli cenné vedomosti a schopnosti, ktoré sú nevyhnutné pre efektívnu aplikáciu stylometrických metód v našej bakalárskej práci.

V priebehu 4. až 5. týždňa sme sa sústredili na dôkladnú analýzu strojového učenia a rôznych algoritmov. Rozoberali sme klúčové algoritmy, ako sú podporné vektorové stroje, náhodné lesy, rozhodovacie stromy, logickú regresiu a naivný bayesov algoritmus. Naše zistenia sú podstatné pre následné fázy projektu, kde plánujeme tieto algoritmy implementovať a testovať na reálnych dátových sadách.

Počas 6. až 8. týždňa sme sa intenzívne venovali analýze existujúcich riešení profilácie autorstva, s osobitným zameraním na identifikáciu veku a pohlavia autora. Toto obdobie bolo významné pre naše porozumenie súčasného stavu v oblasti a pre identifikáciu kľúčových trendov a metodík v tejto špecializovanej doméne. Skúmali sme široké spektrum štúdií a aplikácií, ktoré využívajú rôzne algoritmy strojového učenia a analytické nástroje na extrakciu a interpretáciu štylistických prvkov textu.

Napokon v 9. a 10. týždňi sme si vytvorili zhrnutie analýzy, teda sumarizáciu všetkých získaných informácií a poznatkov z predchádzajúcich fáz štúdia. Na základe zhromaždených údajov a poznatkov sme sa venovali formulácii hypotézy, ktorá by mohla viesť k efektívnejšiemu rozpoznávaniu veku a pohlavia autorov v rámci stylometrických štúdií.

V poslednej časti semestra sme sa sústredili na identifikovanie a riešenie potenciálnych nedostatkov v našej práci, ktoré sme objavili počas sérii konzultačných stretnutí. Táto fáza zahŕňala dôsledné preskúmanie a úpravu nášho dokumentu s cieľom zabezpečiť aby náš záverečný výstup analýzy bol nielen komplexný, ale aj presný.

A.3. Plán práce na letný semester

Plán práce na letný semester sme rozdelili na 4 častí.

Týždeň	Plán práce pre dané obdobie
1 - 2	Extrakcia stylometrických črt
3 - 7	Dokončenie implementácie
8- 11	Testovanie
12 - 13	Dokončenie dokumentu

Tabuľka A.2: Plán práce na letný semester

A.4. Zhodnotenie plánu práce na letný semester

V úvodných dvoch týždňoch sme úspešne extrahovali stylometrické črty, čo zahŕňalo analýzu sentimentu slov a kategorizáciu slov podľa ich sentimentu na pozitívne, negatívne a neutrálne. Použili sme nástroje ako napríklad SentimentIntensityAnalyzer a jazykový model spaCy na identifikáciu entít v texte. Táto fáza bola nevyhnutná pre ďalšie kroky nášho projektu .

V rozmedzí 3. až 7. týždňa sme sa sústredili na dokončenie implementácie. Zahŕňalo to vývoj modelov strojového učenia, ktoré využívali extrahované stylometrické črty. Tieto modely boli optimalizované a pripravované na testovanie. Naša implementácia bola zameraná na zlepšenie presnosti profilácie autorov podľa veku a pohlavia.

V 8. až 11. týždni sme sa zamerali na testovanie zamerané na vyhodnotenie výkonnosti našich modelov. Použili sme klasifikačné metriky ako presnosť, precíznosť, senzitivita a F1 skóre na vyhodnotenie výsledkov. Výsledky testov nám ukázali, že naše modely dosahujú uspokojivú úroveň presnosti a dokážu efektívne rozpoznávať vekové a rodové charakteristiky autorov .

V záverečných týždňoch sme dokončili dokumentáciu našej práce. To zahŕňalo detailné opísanie našich metód, výsledkov a záverov, ktoré sme získali počas letného semestra. Naša dokumentácia bola prehľadná a podrobná, čím poskytovala jasný obraz o našom postupe a zisteniach .

Celkovo sme počas letného semestra splnili všetky plánované úlohy a dosiahli sme významný pokrok v našej bakalárskej práci. Výsledky testovania ukázali, že naše metódy sú efektívne a môžu prispieť k rozvoju nástrojov na profilovanie autorstva v digitálnej ére.

Príloha B

Technická dokumentácia

V tejto prílohe sa zaoberáme technickou dokumentáciou, ktorá detailne popisuje postup inštalácie a spustenia našej implementácie v lokálnom prostredí (localhost). Okrem toho sa budeme venovať aj knižniciam a závislostiam, ktoré boli použité pri vývoji našej aplikácie.

B.1. Systémové požiadavky

- **Minimálne hardvérové požiadavky, ktoré odporúčame pre správne spustenie implementácied/testovaní (Jupyter Notebooks):**
  - CPU: Min. 2.0 GHz
  - RAM: Min. 8 GB
  - Diskový priestor: Min. 500 MB voľného miesta
- **Softvérové požiadavky:**
  - Operačný systém: Windows, macOS, alebo Linux
  - Python: Verzia 3.8 alebo vyššia

B.2. Inštalácia

Postup inštalácie všetkých potrebných komponentov na lokálnom prostredí:

1. **Inštalácia Pythonu:**

(a) Stiahnite si inštalačný balík Pythonu(Verzia 3.8+) z oficiálnej webovej stránky python.org.

(b) Počas inštalácie zaškrtnite možnosť *Add Python to PATH*.

(c) Dokončite inštaláciu podľa pokynov inštalátora.

2. **Klonovanie repozitára (voliteľné ak máme k dipozíícii prístup k digitálnej časti BP :**

```
git clone https://github.com/DanielaPaluchova/
BP_DanielaPaluchova
cd BP_DanielaPaluchova
```

3. **Vytvorenie a aktivácia virtuálneho prostredia:**

```
python -m venv venv
Pre macOS alebo Linux: source venv/bin/activate
Pre Windows: venv\Scripts\activate
```

4. **Inštalácia závislostí:**

```
pip install -r requirements.txt
```

B.3. Spustenie aplikácie

Detailný popis, ako spustiť aplikáciu:

1. **Nastavenie environmentálnych premenných:** Uistite sa, že máte správne nastavené environmentálne premenné pre konfiguráciu aplikácie.

2. **Spustenie Jupyter Notebooku:**

```
jupyter notebook
```

3. **Načítanie notebookov:** Po spustení Jupyter Notebooku sa automaticky otvorí webový prehliadač. V ňom prejdite na notebooky obsahujúce implementácie a testovania (v tomto prípade je to rovnaké) a otvorte ich. K dispozícii je veľké množstvo notebookov, pričom každý je podrobne opísaný a z jeho názvu obvykle vyplýva, o aký druh implementácie sa jedná. Ukážka:

   - **Testovanie:** `vekova_kategoria_fs_metoda.ipynb`

B.4. Použité knižnice a závislosti

V tejto kapitole sa venujeme použitým knižniciam a závislostiam. Krátko opisujeme ich účel a uvádzame, ako sme ich využili v našom projekte.b

- **pip**
  - Systém na správu balíkov pre Python, ktorý umožňuje inštaláciu a správu softvérových balíkov.
- **Python**
  - Vysokoúrovňový programovací jazyk známy svojou čitateľnosťou a jednoduchou syntaxou, ktorý podporuje rôzne programovacie paradigmy.
- **Pandas**
  - Knižnica na manipuláciu a analýzu dát, ktorá poskytuje dátové štruktúry a nástroje na jednoduchú prácu s relačnými alebo označenými dátami.
- **NumPy**
  - Knižnica pre prácu s n-dimenzionálnymi poľami a maticami, ktorá poskytuje širokú škálu matematických funkcií na vysokoúrovňové operácie.
- **Matplotlib**
  - Knižnica na tvorbu statických, animovaných a interaktívnych vizualizácií v Pythone. Bola použitá na tvorbu grafov a vizualizácií pri analýze veku a pohlavia.

- **Seaborn**
  - Knižnica na vizualizáciu dát založená na Matplotlib, ktorá ponúka atraktívne a informatívne štatistické grafy. Použitá na detailnú vizualizáciu štatistík veku a pohlavia.
- **SciPy**
  - Knižnica na vedecké a technické výpočty, ktorá rozširuje funkčnosť NumPy o pokročilé matematické, technické a vedecké funkcie.
- **jsonlines**
  - Knižnica na čítanie a zapisovanie dát vo formáte JSON Lines, kde každá línia predstavuje samostatný JSON objekt.
- **BeautifulSoup4**
  - Knižnica na analýzu (parsovanie) HTML a XML dokumentov, ktorá uľahčuje extrahovanie dát z webových stránok.
- **re**
  - Modul na prácu s regulárnymi výrazmi, ktorý umožňuje vyhľadávanie a manipuláciu s textovými reťazcami.
- **nltk**
  - Knižnica pre spracovanie prirodzeného jazyka, ktorá poskytuje nástroje na analýzu textu, tokenizáciu, označovanie častí reči, a mnoho ďalších úloh v NLP. Použitá na analýzu textových údajov pri profilácii veku a pohlavia.
- **datetime**
  - Modul na prácu s dátumom a časom, ktorý umožňuje vytváranie, manipuláciu a formátovanie dátumových a časových objektov.
- **emoji**
  - Knižnica na prácu s emodži v textových reťazcoch, ktorá umožňuje ich pridávanie, vyhľadávanie a manipuláciu.

- **sklearn**
  - Známa ako scikit-learn, je knižnica na strojové učenie, ktorá ponúka jednoduché a efektívne nástroje na analýzu a modelovanie dát. Použitá na modelovanie a predikciu veku a pohlavia.
- **spaCy**
  - Knižnica pre spracovanie prirodzeného jazyka, ktorá je navrhnutá pre výkon a produkčné použitie, ponúkajúca pokročilé NLP funkcie ako tokenizáciu, označovanie častí reči, a závislostnú analýzu.
- **langdetect**
  - Knižnica na detekciu jazyka textu, ktorá je založená na Google Ngram modeli a podporuje viacero jazykov.
- **Jupyter**
  - Interaktívne prostredie na tvorbu a zdieľanie dokumentov, ktoré kombinujú kód, rovnice, vizualizácie a text. Použité na vývoj a testovanie modelov pre profiláciu veku a pohlavia.
- **scikit-learn**
  - Knižnica na strojové učenie, ktorá poskytuje jednoduché a efektívne nástroje pre prediktívne dátové analýzy a modelovanie. Použitá na modelovanie a predikciu veku a pohlavia.

Detailný zoznam závislostí nájdete v súbore `requirements.txt`.

B.5. Problémy a ich riešenie

Bežné problémy, na ktoré môže užívateľ naraziť počas inštalácie alebo používania aplikácie:

- **Problém:** Inštalácia závislostí zlyhá.
  - **Riešenie:** Skontrolujte, či máte správne nastavené virtuálne prostredie a že používate správnu verziu Pythonu.

- **Problém:** Aplikácia nenačíta vstupný .json súbor(dataset).
  - **Riešenie:** Uistite sa, že cesta k súboru je správna a súbor existuje.

Príloha C

Zoznam extrahovaných črt

V tejto prílohe prikladáme všetky nami extrahované čty z textu, ktoré sme využili pri implementáciach.

C.1. Lexikálne črty

- Priemerný počet emoji na tweet
- Priemerný počet zmienok na tweet
- Priemerný počet hashtagov na tweet
- Priemerný počet cifier na tweet
- Priemerný počet veľkých písmen na tweet
- Priemerný počet malých písmen na tweet
- Priemerný počet nealfabetických znakov na tweet
- Priemerný počet uvodzoviek na tweet
- Priemerný počet apostrofov na tweet
- Priemerný počet interpunkčných značiek na tweet
- Priemerný počet viacnásobných interpunkčných značiek na tweet
- Priemerný počet zastavovacích slov na tweet

- Priemerná dĺžka slova
- Priemerný počet slov na vetu
- Priemerný počet znakov na vetu
- Priemerný počet viet na tweet
- Priemerný počet slov na tweet
- Priemerný počet znakov na tweet
- Priemerný počet násobne opakovaných znakov na tweet
- Veľkosť slovnej zásoby

C.2. Syntaktické črty

- Percento oznamovacích viet
- Percento otázok
- Percento rozkazovacích viet
- Percento neštandardných ukončení viet
- Priemerný počet čiarok na tweet
- Priemerný počet podstatných mien na tweet
- Priemerný počet prídavných mien na tweet
- Priemerný počet slovies na tweet
- Priemerný počet zámen na tweet
- Priemerný počet čísloviek na tweet
- Priemerný počet príslovek na tweet
- Priemerný počet predložiek na tweet

- Priemerný počet spojok na tweet
- Priemerný počet častíc na tweet
- Priemerný počet citoslovci na tweet
- Priemerný počet minulých časov na tweet
- Priemerný počet prítomných časov na tweet
- Priemerný počet podmetov na tweet
- Priemerný počet priamych predmetov na tweet
- Priemerný počet príslovných určení na tweet
- Priemerný počet prívlastkov na tweet
- Priemerný počet pomocných slovies na tweet

C.3. Obsahové črty

- Priemerný počet pozitívnych slov na tweet
- Priemerný počet negatívnych slov na tweet
- Priemerný počet neutrálnych slov na tweet
- Priemerný počet entít na tweet
- Priemerné skóre sentimentu
- Priemerný počet negácií na tweet
- Priemerný počet kognitívnych slov na tweet
- Priemerný počet zmyslových slov na tweet

Príloha D

Testovanie

V tejto prílohe sa nachadzáju tabuľky výstupov z nášho testovania.

D.1. Pohlavie testovanie

Celkovo sme riešili 3 druhy testovania pohlavia.

D.1.1 Testovanie 1

Varianta datasetu	**accuracy_test**	**precision_test**	**recall_test**	**f1_test**
dataset bez ošetrenia outlierov a škálovania	0.701716	0.701716	1	0.824683
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.759804	0.750508	0.985354	0.851987
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.701716	0.701716	1	0.824683
dataset bez ošetrenia outlierov a škálovania s word frequency	0.74853	0.739621	0.990411	0.846777
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.701716	0.701716	1	0.824683

Varianta datasetu	**accuracy_test**	**precision_test**	**recall_test**	**f1_test**
dataset s ošetrením outlierov	0.701716	0.701716	1	0.824683
dataset s ošetrením outlierov a s ngramami	0.765931	0.755441	0.985526	0.855234
dataset s ošetrením outlierov a s tfidf ngramami	0.701716	0.701716	1	0.824683
dataset s ošetrením outlierov a s word frequency	0.758946	0.748533	0.98867	0.851958
dataset s ošetrením outlierov a s tfidf word frequency	0.701716	0.701716	1	0.824683
dataset so skalovaním	0.701716	0.701716	1	0.824683
dataset so skalovaním a s ngramami	0.746446	0.738734	0.988305	0.845425
dataset so skalovaním a s tfidf ngramami	0.701716	0.701716	1	0.824683
dataset so skalovaním a s word frequency	0.733946	0.727561	0.992679	0.839629
dataset so skalovaním a s tfidf word frequency	0.701716	0.701716	1	0.824683
dataset s ošetrením outlierov a skalovaním	0.701716	0.701716	1	0.824683
dataset s ošetrením outlierov a skalovaním a s ngramami	0.765931	0.755441	0.985526	0.855234
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.701716	0.701716	1	0.824683
dataset s ošetrením outlierov a skalovaním a s word frequency	0.758946	0.748533	0.98867	0.851958

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.701716	0.701716	1	0.824683

Tabuľka D.1: Pohlavie - Testovanie 1 SVC

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.761397167	0.7678825	0.946120167	0.847660333
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.845098	0.834339667	0.972448667	0.898053667
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.847794333	0.8414485	0.965311167	0.898991167
dataset bez ošetrenia outlierov a škálovania s word frequency	0.848407	0.837660167	0.972669333	0.900049833
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.849632333	0.840518333	0.970044167	0.900529333
dataset s ošetrením outlierov	0.909068833	0.9022295	0.976282	0.937752
dataset s ošetrením outlierov a s ngramami	0.874632333	0.860659667	0.980120833	0.916454833
dataset s ošetrením outlierov a s tfidf ngramami	0.884068667	0.8722795	0.978548833	0.9222055
dataset s ošetrením outlierov a s word frequency	0.908333333	0.8940495	0.9864105	0.937895
dataset s ošetrením outlierov a s tfidf word frequency	0.897303833	0.883355833	0.983779833	0.930783667

Varianta datasetu	**accuracy_test**	**precision_test**	**recall_test**	**f1_test**
dataset so skalovaním	0.765318667	0.7682685	0.953303833	0.850751167
dataset so skalovaním a s ngramami	0.8460785	0.836445333	0.970696667	0.898481667
dataset so skalovaním a s tfidf ngramami	0.852083167	0.8432245	0.969635167	0.901945667
dataset so skalovaním a s word frequency	0.848897167	0.8377705	0.973368	0.900405
dataset so skalovaním a s tfidf word frequency	0.848161833	0.838084167	0.9713715	0.8997525
dataset s ošetrením outlierov a skalovaním	0.9088235	0.905316667	0.971732833	0.937329333
dataset s ošetrením outlierov a skalovaním a s ngramami	0.877206	0.862561167	0.981512167	0.918132333
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.887867667	0.876162167	0.978894833	0.924551667
dataset s ošetrením outlierov a skalovaním a s word frequency	0.906004667	0.892996667	0.984131167	0.936276667
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.897426333	0.883342333	0.983945333	0.930859

Tabuľka D.2: Pohlavie - Testovanie 1 RF

Varianta datasetu	**accuracy_test**	**precision_test**	**recall_test**	**f1_test**
dataset bez ošetrenia outlierov a škálovania	0.768015	0.781823	0.928847	0.848458

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.899387	0.907949	0.953413	0.930055
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.888603	0.898565	0.948346	0.922742
dataset bez ošetrenia outlierov a škálovania s word frequency	0.885539	0.897463	0.945056	0.920561
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.886765	0.899251	0.944498	0.921271
dataset s ošetrením outlierov	0.936152	0.933733	0.978564	0.955582
dataset s ošetrením outlierov a s ngramami	0.961152	0.960321	0.985337	0.972664
dataset s ošetrením outlierov a s tfidf ngramami	0.955024	0.95391	0.983422	0.968436
dataset s ošetrením outlierov a s word frequency	0.960784	0.960306	0.984831	0.972405
dataset s ošetrením outlierov a s tfidf word frequency	0.956985	0.954955	0.985188	0.969829
dataset so skalovaním	0.768505	0.782366	0.928687	0.849154
dataset so skalovaním a s ngramami	0.899755	0.908133	0.953757	0.930314
dataset so skalovaním a s tfidf ngramami	0.888603	0.898697	0.948173	0.92273
dataset so skalovaním a s word frequency	0.885539	0.897466	0.945056	0.920564
dataset so skalovaním a s tfidf word frequency	0.88701	0.899284	0.944849	0.921455

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a skalovaním	0.936152	0.933733	0.978564	0.955582
dataset s ošetrením outlierov a skalovaním a s ngramami	0.961274	0.960486	0.985337	0.972749
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.954902	0.95375	0.983422	0.968353
dataset s ošetrením outlierov a skalovaním a s word frequency	0.960784	0.960306	0.984831	0.972405
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.956985	0.954955	0.985188	0.969829

Tabuľka D.3: Pohlavie - Testovanie 1 GBM

D.1.2 Testovanie 2

Pri RFE nám vyšla takáto najlepšia kombinácia vlastnosti:

- **SVC pre dataset bez ošetrenia outlierov a škálovania :** priemerný počet emoji na tweet, priemerný počet URL adries na tweet, priemerný počet úvodzoviek na tweet, priemerný počet viacnásobnej interpunkcie za sebou na tweet, priemerná dĺžka slova, priemerný počet slov na tweet, priemerný počet opakovaných znakov na tweet, priemerný počet prídavných mien na tweet, priemerný počet slovies na tweet, priemerný počet zámen na tweet, priemerný počet čísloviek na tweet, priemerný počet prísloviek na tweet, priemerný počet častíc na tweet, priemerný počet citosloviec na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet podmetov na tweet, priemerný počet priamych predmetov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet prívlastkov na tweet, priemerný počet pomocných slovies na tweet, priemerný počet negatívnych slov na tweet, priemerné sentimentové skóre, priemerný počet negácií na tweet, priemerný počet kognitívnych slov na

tweet

- **SVC pre dataset s ošetrením outlierov :** priemerný počet emoji na tweet, priemerný počet URL adries na tweet, priemerný počet apostrofov na tweet, priemerná dĺžka slova, priemerný počet opakovaných znakov na tweet, percentuálny podiel neštandardných zakončení viet na tweet, priemerný počet čiarok na tweet, priemerný počet zámen na tweet, priemerný počet číslovieк na tweet, priemerný počet prísloviek na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, priemerný počet citosloviec na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet vyjadrených podmetov na tweet, priemerný počet priamych predmetov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet prívlastkov na tweet, priemerný počet pomocných slovies na tweet, priemerný počet entít na tweet, priemerné sentimentové skóre, priemerný počet negácií na tweet, priemerný počet kognitívnych slov na tweet, priemerný počet zmyslových slov na tweet

- **SVC pre dataset so skalovaním :** priemerný počet emoji na tweet, priemerný počet URL adries na tweet, priemerný počet apostrofov na tweet, priemerná dĺžka slova, priemerný počet opakovaných znakov na tweet, percentuálny podiel otázok na tweet, percentuálny podiel rozkazovacích viet na tweet, percentuálny podiel neštandardných ukončení viet na tweet, priemerný počet čiarok na tweet, priemerný počet podstatných mien na tweet, priemerný počet prídavných mien na tweet, priemerný počet slovies na tweet, priemerný počet zámen na tweet, priemerný počet číslovieк na tweet, priemerný počet prísloviek na tweet, priemerný počet predložiek na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, priemerný počet citosloviec na tweet, priemerný počet prítomných časov na tweet, priemerný počet prívlastkov na tweet, priemerný počet entít na tweet, priemerný počet negácií na tweet, priemerný počet kognitívnych slov na tweet, priemerný počet zmyslových slov na tweet

- **SVC pre dataset s ošetrením outlierov a skalovaním :** priemerný počet emoji na tweet, priemerný počet URL adries na tweet, priemerný počet apostrofov na tweet, priemerná dĺžka slova, priemerný počet opakovaných znakov na tweet, percentuálny podiel neštandardných ukončení viet na tweet, priemerný počet

čiarok na tweet, priemerný počet zámen na tweet, priemerný počet čísloviek na tweet, priemerný počet prísloviek na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, priemerný počet citosloviec na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet podmetov na tweet, priemerný počet priamych predmetov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet prívlastkov na tweet, priemerný počet pomocných slovies na tweet, priemerný počet entít na tweet, priemerné sentimentové skóre, priemerný počet negácií na tweet, priemerný počet kognitívnych slov na tweet, priemerný počet zmyslových slov na tweet

- **RF pre dataset bez ošetrenia outlierov a škálovania :** priemerný počet emoji na tweet, priemerný počet URL adries na tweet, priemerný počet hashtagov na tweet, priemerný počet číslic na tweet, priemerný počet veľkých písmen na tweet, priemerný počet neabecedných znakov na tweet, priemerný počet interpunkčných znamienok na tweet, priemerný počet viacnásobnej interpunkcie za sebou na tweet, priemerná dĺžka slova, priemerný počet opakovaných znakov na tweet, percentuálny podiel oznamovacích viet na tweet, percentuálny podiel rozkazovacích viet na tweet, percentuálny podiel ostatných koncov na tweet, priemerný počet zámen na tweet, priemerný počet čísloviek na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, priemerný počet citosloviec na tweet, priemerný počet prítomných časov na tweet, priemerný počet negatívnych slov na tweet, priemerný počet entít na tweet, priemerné sentimentové skóre, priemerný počet negácií na tweet, priemerný počet kognitívnych slov na tweet, priemerný počet zmyslových slov na tweet
- **RF pre dataset s ošetrením outlierov :** priemerný počet emoji na tweet, priemerný počet URL adries na tweet, priemerný počet neabecedných znakov na tweet, priemerná dĺžka slova, priemerný počet opakovaných znakov na tweet, percentuálny podiel oznamovacích viet na tweet, percentuálny podiel otázok na tweet, percentuálny podiel rozkazovacích viet na tweet, percentuálny podiel ostatných koncov na tweet, priemerný počet čiarok na tweet, priemerný počet podstatných mien na tweet, priemerný počet prídavných mien na tweet, priemerný počet slovies na tweet, priemerný počet zámen na tweet, priemerný počet číslo-

viek na tweet, priemerný počet prísloviek na tweet, priemerný počet predložiek na tweet, priemerný počet častíc na tweet, priemerný počet citosloviec na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet podmetov na tweet, priemerný počet priamych predmetov na tweet, priemerný počet príslovkových modifikátorov na tweet, priemerný počet prívlastkov na tweet

- **RF pre dataset so skalovaním :** priemerný počet emoji na tweet, priemerný počet URL adries na tweet, priemerný počet spomenutí (@mentions) na tweet, priemerný počet hashtagov na tweet, priemerný počet číslic na tweet, priemerný počet veľkých písmen na tweet, priemerný počet neabecedných znakov na tweet, priemerný počet interpunkčných znamienok na tweet, priemerný počet viacnásobnej interpunkcie za sebou na tweet, priemerná dĺžka slova, priemerný počet slov na vetu, priemerný počet opakovaných znakov na tweet, percentuálny podiel oznamovacích viet na tweet, percentuálny podiel rozkazovacích viet na tweet, percentuálny podiel ostatných koncov na tweet, priemerný počet zámen na tweet, priemerný počet čísloviek na tweet, priemerný počet častíc na tweet, priemerný počet citosloviec na tweet, priemerný počet prítomných časov na tweet, priemerný počet príslovkových modifikátorov na tweet, priemerný počet negatívnych slov na tweet, priemerný počet entít na tweet, priemerné sentimentové skóre, priemerný počet negácií na tweet

- **RF pre dataset s ošetrením outlierov a skalovaním :** priemerný počet emoji na tweet, priemerný počet URL adries na tweet, priemerný počet neabecedných znakov na tweet, priemerná dĺžka slova, priemerný počet opakovaných znakov na tweet, percentuálny podiel oznamovacích viet na tweet, percentuálny podiel otázok na tweet, percentuálny podiel rozkazovacích viet na tweet, percentuálny podiel ostatných koncov na tweet, priemerný počet čiarok na tweet, priemerný počet podstatných mien na tweet, priemerný počet prídavných mien na tweet, priemerný počet slovies na tweet, priemerný počet zámen na tweet, priemerný počet čísloviek na tweet, priemerný počet prísloviek na tweet, priemerný počet predložiek na tweet, priemerný počet častíc na tweet, priemerný počet citosloviec na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet podmetov na tweet, priemerný počet priamych

predmetov na tweet, priemerný počet príslovkových modifikátorov na tweet, priemerný počet prívlastkov na tweet

- **GBM pre dataset bez ošetrenia outlierov a škálovania :** priemerný počet emoji na tweet, priemerný počet URL adries na tweet, priemerný počet hashtagov na tweet, priemerný počet číslic na tweet, priemerný počet malých písmen na tweet, priemerný počet neabecedných znakov na tweet, priemerný počet interpunkčných znamienok na tweet, priemerná dĺžka slova, priemerný počet slov na vetu, priemerný počet slov na tweet, priemerný počet opakovaných znakov na tweet, percentuálny podiel oznamovacích viet na tweet, percentuálny podiel otázok na tweet, percentuálny podiel rozkazovacích viet na tweet, percentuálny podiel ostatných koncov na tweet, priemerný počet zámen na tweet, priemerný počet čísloviek na tweet, priemerný počet častíc na tweet, priemerný počet citosloviec na tweet, priemerný počet prítomných časov na tweet, priemerný počet pomocných slovies na tweet, priemerný počet negatívnych slov na tweet, priemerný počet entít na tweet, priemerné sentimentové skóre, priemerný počet negácií na tweet
- **GBM pre dataset s ošetrením outlierov :** priemerný počet emoji na tweet, priemerný počet URL adries na tweet, priemerný počet stop slov na tweet, priemerná dĺžka slova, priemerný počet znakov na vetu, priemerný počet opakovaných znakov na tweet, percentuálny podiel oznamovacích viet na tweet, percentuálny podiel otázok na tweet, percentuálny podiel rozkazovacích viet na tweet, percentuálny podiel ostatných koncov na tweet, priemerný počet čiarok na tweet, priemerný počet podstatných mien na tweet, priemerný počet slovies na tweet, priemerný počet zámen na tweet, priemerný počet čísloviek na tweet, priemerný počet prísloviek na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, priemerný počet citosloviec na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet podmetov na tweet, priemerný počet priamych predmetov na tweet, priemerný počet príslovkových modifikátorov na tweet, priemerný počet prívlastkov na tweet
- **GBM pre dataset so skalovaním :** priemerný počet emoji na tweet, priemerný počet URL adries na tweet, priemerný počet hashtagov na tweet, priemerný počet číslic na tweet, priemerný počet malých písmen na tweet, priemerný počet in-

terpunkčných znamienok na tweet, priemerná dĺžka slova, priemerný počet slov na vetu, priemerný počet slov na tweet, priemerný počet opakovaných znakov na tweet, percentuálny podiel oznamovacích viet na tweet, percentuálny podiel otázok na tweet, percentuálny podiel rozkazovacích viet na tweet, percentuálny podiel ostatných koncov na tweet, priemerný počet zámen na tweet, priemerný počet čísloviek na tweet, priemerný počet častíc na tweet, priemerný počet citosloviec na tweet

Varianta datasetu	**accuracy_test**	**precision_test**	**recall_test**	**f1_test**
dataset bez ošetrenia outlierov a škálovania	0.703431	0.703431	1	0.825854
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.840441	0.835985	0.962086	0.894518
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.705392	0.705011	0.999307	0.826708
dataset bez ošetrenia outlierov a škálovania s word frequency	0.830392	0.82878	0.956858	0.888086
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.705147	0.704838	0.999307	0.826589
dataset s ošetrením outlierov	0.726716	0.725469	0.985741	0.835488
dataset s ošetrením outlierov a s ngramami	0.840196	0.835725	0.962086	0.894372
dataset s ošetrením outlierov a s tfidf ngramami	0.748039	0.743699	0.984013	0.846445
dataset s ošetrením outlierov a s word frequency	0.830637	0.829026	0.956858	0.888229
dataset s ošetrením outlierov a s tfidf word frequency	0.747059	0.742881	0.984024	0.845921

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset so skalovaním	0.703922	0.703784	1	0.826093
dataset so skalovaním a s ngramami	0.840196	0.835725	0.962086	0.894372
dataset so skalovaním a s tfidf ngramami	0.759804	0.753306	0.979464	0.851535
dataset so skalovaním a s word frequency	0.830392	0.82878	0.956858	0.888086
dataset so skalovaním a s tfidf word frequency	0.759069	0.752326	0.98051	0.851288
dataset s ošetrením outlierov a skalovaním	0.726716	0.725469	0.985741	0.835488
dataset s ošetrením outlierov a skalovaním a s ngramami	0.840196	0.835725	0.962086	0.894372
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.748039	0.743699	0.984013	0.846445
dataset s ošetrením outlierov a skalovaním a s word frequency	0.830637	0.829026	0.956858	0.888229
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.747059	0.742881	0.984024	0.845921

Tabuľka D.4: Pohlavie - Testovanie 2 RFE SVC

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.771323	0.778261	0.944068	0.853102

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.843137	0.83369	0.970781	0.896946
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.851225	0.844699	0.966247	0.901315
dataset bez ošetrenia outlierov a škálovania s word frequency	0.845588	0.835463	0.97219	0.898542
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.846569	0.838876	0.967997	0.898722
dataset s ošetrením outlierov	0.908333	0.906487	0.969725	0.937015
dataset s ošetrením outlierov a s ngramami	0.869853	0.855813	0.98018	0.913725
dataset s ošetrením outlierov a s tfidf ngramami	0.887745	0.875669	0.979463	0.924634
dataset s ošetrením outlierov a s word frequency	0.89951	0.88661	0.983007	0.932242
dataset s ošetrením outlierov a s tfidf word frequency	0.890932	0.877853	0.981579	0.92676
dataset so skalovaním	0.769118	0.777421	0.941558	0.851541
dataset so skalovaním a s ngramami	0.841912	0.830827	0.973554	0.896491
dataset so skalovaním a s tfidf ngramami	0.849265	0.842238	0.966994	0.900232
dataset so skalovaním a s word frequency	0.843628	0.833352	0.972145	0.897355
dataset so skalovaním a s tfidf word frequency	0.854902	0.84458	0.972893	0.904122

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a skalovaním	0.910784	0.909181	0.970077	0.938612
dataset s ošetrením outlierov a skalovaním a s ngramami	0.870098	0.854753	0.982304	0.914056
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.890686	0.879199	0.979135	0.926445
dataset s ošetrením outlierov a skalovaním a s word frequency	0.902941	0.889246	0.984682	0.934472
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.899265	0.885105	0.984681	0.932167

Tabuľka D.5: Pohlavie - Testovanie 2 RFE RF

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.769118	0.78625	0.923146	0.849036
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.898039	0.90692	0.953048	0.929322
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.88701	0.898974	0.94572	0.9217
dataset bez ošetrenia outlierov a škálovania s word frequency	0.882108	0.894337	0.943935	0.918438
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.884068	0.898042	0.942199	0.919553
dataset s ošetrením outlierov	0.935784	0.930656	0.981951	0.955561

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a s ngramami	0.959314	0.95861	0.984648	0.971453
dataset s ošetrením outlierov a s tfidf ngramami	0.956863	0.95566	0.984351	0.969787
dataset s ošetrením outlierov a s word frequency	0.959314	0.957344	0.986068	0.971483
dataset s ošetrením outlierov a s tfidf word frequency	0.957843	0.956969	0.984352	0.970457
dataset so skalovaním	0.768872	0.783933	0.927316	0.849464
dataset so skalovaním a s ngramami	0.896569	0.905158	0.953037	0.92838
dataset so skalovaním a s tfidf ngramami	0.888971	0.9003	0.947084	0.923045
dataset so skalovaním a s word frequency	0.880392	0.893852	0.941877	0.91719
dataset so skalovaním a s tfidf word frequency	0.884314	0.898616	0.941867	0.919686
dataset s ošetrením outlierov a skalovaním	0.936029	0.930965	0.981951	0.955724
dataset s ošetrením outlierov a skalovaním a s ngramami	0.959314	0.95861	0.984648	0.971453
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.956863	0.95566	0.984351	0.969787
dataset s ošetrením outlierov a skalovaním a s word frequency	0.959559	0.957673	0.986068	0.971652

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.957843	0.956969	0.984352	0.970457

Tabuľka D.6: Pohlavie - Testovanie 2 RFE GBM

Najlepšie kombinácie vlastnoti pre FS:

- **SVC pre dataset bez ošetrenia outlierov a škálovania :** priemerný počet opakovaných znakov na tweet, priemerné sentimentové skóre, priemerný počet kognitívnych slov na tweet
- **SVC pre dataset s ošetrením outlierov :** percentuálny podiel rozkazovacích viet na tweet, priemerný počet neabecedných znakov na tweet
- **SVC pre dataset so skalovaním :** priemerný počet opakovaných znakov na tweet, priemerný počet častíc na tweet, priemerné sentimentové skóre, priemerný počet citosloviec na tweet, priemerný počet predložiek na tweet
- **SVC pre dataset s ošetrením outlierov a skalovaním :** percentuálny podiel rozkazovacích viet na tweet, priemerný počet neabecedných znakov na tweet
- **RF pre dataset bez ošetrenia outlierov a škálovania :** priemerný počet citosloviec na tweet
- **RF pre dataset s ošetrením outlierov :** priemerný počet citosloviec na tweet, percentuálny podiel rozkazovacích viet na tweet, priemerný počet čísloviek na tweet, priemerný počet prítomných časov na tweet, percentuálny podiel otázok na tweet, priemerný počet zámen na tweet, priemerný počet prísloviek na tweet, priemerný počet priamych predmetov na tweet, priemerný počet prídavných mien na tweet, priemerný počet častíc na tweet, priemerný počet stop slov na tweet, priemerná dĺžka slova, priemerný počet minulých časov na tweet
- **RF pre dataset so skalovaním :** priemerný počet citosloviec na tweet
- **RF pre dataset s ošetrením outlierov a skalovaním :** priemerný počet citosloviec na tweet, percentuálny podiel rozkazovacích viet na tweet, priemerný počet

priamych predmetov na tweet, priemerný počet podstatných mien na tweet, priemerný počet zámen na tweet, priemerný počet čísloviek na tweet, priemerný počet entít na tweet

- **GBM pre dataset bez ošetrenia outlierov a škálovania :** priemerný počet opakovaných znakov na tweet, priemerný počet neabecedných znakov na tweet, priemerný počet neutrálnych slov na tweet, priemerné sentimentové skóre, priemerný počet prítomných časov na tweet, priemerná dĺžka slova, priemerný počet pomocných slovies na tweet, percentuálny podiel ostatných koncov na tweet, priemerný počet emoji na tweet, priemerný počet negácií na tweet, priemerný počet príslovkových modifikátorov na tweet

- **GBM pre dataset s ošetrením outlierov :** percentuálny podiel rozkazovacích viet na tweet, priemerný počet čísloviek na tweet, priemerný počet podmetov na tweet, priemerný počet častíc na tweet, percentuálny podiel otázok na tweet, priemerný počet zámen na tweet, percentuálny podiel ostatných koncov na tweet, priemerný počet prísloviek na tweet, priemerný počet čiarok na tweet, priemerný počet apostrofov na tweet, priemerný počet predložiek na tweet, priemerný počet viet na tweet

- **GBM pre dataset so skalovaním :** priemerný počet opakovaných znakov na tweet, priemerný počet neabecedných znakov na tweet, priemerný počet entít na tweet, priemerný počet emoji na tweet, priemerný počet pozitívnych slov na tweet, priemerný počet pomocných slovies na tweet

- **GBM pre dataset s ošetrením outlierov a skalovaním :** percentuálny podiel rozkazovacích viet na tweet, priemerný počet čísloviek na tweet, priemerný počet podmetov na tweet, priemerný počet častíc na tweet, percentuálny podiel otázok na tweet, priemerný počet zámen na tweet, percentuálny podiel ostatných koncov na tweet, priemerný počet prísloviek na tweet, priemerný počet čiarok na tweet, priemerný počet apostrofov na tweet, priemerný počet predložiek na tweet, priemerný počet viet na tweet

Varianta datasetu	**accuracy_test**	**precision_test**	**recall_test**	**f1_test**
dataset bez ošetrenia outlierov a škálovania	0.720098	0.719116	0.985359	0.831334
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.844363	0.838272	0.963568	0.896537
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.875	0.875237	0.95802	0.91474
dataset bez ošetrenia outlierov a škálovania s word frequency	0.837745	0.833276	0.960431	0.892305
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.873284	0.871244	0.96118	0.913931
dataset s ošetrením outlierov	0.730882	0.729166	0.98148	0.836266
dataset s ošetrením outlierov a s ngramami	0.845588	0.841256	0.960773	0.897006
dataset s ošetrením outlierov a s tfidf ngramami	0.823284	0.824053	0.962544	0.886059
dataset s ošetrením outlierov a s word frequency	0.840196	0.837298	0.957985	0.893529
dataset s ošetrením outlierov a s tfidf word frequency	0.822059	0.823412	0.961512	0.885212
dataset so skalovaním	0.72451	0.724226	0.980488	0.832893
dataset so skalovaním a s ngramami	0.844363	0.838272	0.963568	0.896537
dataset so skalovaním a s tfidf ngramami	0.875735	0.876082	0.958023	0.915197

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset so skalovaním a s word frequency	0.837745	0.833276	0.960431	0.892305
dataset so skalovaním a s tfidf word frequency	0.872059	0.871532	0.958746	0.912973
dataset s ošetrením outlierov a skalovaním	0.730882	0.729166	0.98148	0.836266
dataset s ošetrením outlierov a skalovaním a s ngramami	0.845588	0.841256	0.960773	0.897006
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.823284	0.824053	0.962544	0.886059
dataset s ošetrením outlierov a skalovaním a s word frequency	0.840196	0.837298	0.957985	0.893529
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.822059	0.823412	0.961512	0.885212

Tabuľka D.7: Pohlavie - Testovanie 2 FS SVC

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.740196	0.765251	0.907669	0.830368
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.846568	0.837253	0.969501	0.898444
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.847059	0.837158	0.970649	0.89885
dataset bez ošetrenia outlierov a škálovania s word frequency	0.852206	0.841379	0.972372	0.902071

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.854657	0.844572	0.971321	0.903428
dataset s ošetrením outlierov	0.908579	0.910671	0.963947	0.936538
dataset s ošetrením outlierov a s ngramami	0.866667	0.852187	0.979704	0.911422
dataset s ošetrením outlierov a s tfidf ngramami	0.864461	0.855519	0.97097	0.909398
dataset s ošetrením outlierov a s word frequency	0.8875	0.870674	0.986013	0.924658
dataset s ošetrením outlierov a s tfidf word frequency	0.867402	0.854811	0.976927	0.911658
dataset so skalovaním	0.691912	0.734007	0.878179	0.799607
dataset so skalovaním a s ngramami	0.846569	0.836629	0.97061	0.898556
dataset so skalovaním a s tfidf ngramami	0.849265	0.840443	0.969251	0.900066
dataset so skalovaním a s word frequency	0.84902	0.83624	0.975535	0.900461
dataset so skalovaním a s tfidf word frequency	0.84853	0.838884	0.9703	0.899684
dataset s ošetrením outlierov a skalovaním	0.901961	0.907697	0.957285	0.931817
dataset s ošetrením outlierov a skalovaním a s ngramami	0.860049	0.846156	0.978336	0.907321
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.861029	0.850202	0.973391	0.907501

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a skalovaním a s word frequency	0.880392	0.866511	0.980387	0.919866
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.871569	0.858269	0.978684	0.914352

Tabuľka D.8: Pohlavie - Testovanie 2 FS RF

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.770343	0.77938	0.93735	0.851088
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.903676	0.91154	0.955204	0.932801
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.890441	0.901923	0.946473	0.923625
dataset bez ošetrenia outlierov a škálovania s word frequency	0.892157	0.904147	0.946484	0.924734
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.891421	0.90276	0.947166	0.924335
dataset s ošetrením outlierov	0.93799	0.934871	0.979709	0.956753
dataset s ošetrením outlierov a s ngramami	0.954902	0.956924	0.979737	0.968166
dataset s ošetrením outlierov a s tfidf ngramami	0.948774	0.950625	0.977611	0.963927
dataset s ošetrením outlierov a s word frequency	0.959559	0.961253	0.981819	0.971414

Varianta datasetu	**accuracy_test**	**precision_test**	**recall_test**	**f1_test**
dataset s ošetrením outlierov a s tfidf word frequency	0.953186	0.952456	0.982196	0.967079
dataset so skalovaním	0.767647	0.775603	0.940203	0.84999
dataset so skalovaním a s ngramami	0.90098	0.910964	0.951711	0.930819
dataset so skalovaním a s tfidf ngramami	0.889216	0.901537	0.945087	0.92273
dataset so skalovaním a s word frequency	0.893627	0.904132	0.948931	0.925873
dataset so skalovaním a s tfidf word frequency	0.892157	0.904194	0.94646	0.924751
dataset s ošetrením outlierov a skalovaním	0.934314	0.932202	0.977227	0.954171
dataset s ošetrením outlierov a skalovaním a s ngramami	0.955392	0.957886	0.979383	0.968491
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.948774	0.95001	0.97832	0.963951
dataset s ošetrením outlierov a skalovaním a s word frequency	0.959069	0.960913	0.981464	0.971064
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.95049	0.951342	0.979359	0.965134

Tabuľka D.9: Pohlavie - Testovanie 2 FS GBM

D.1.3 Testovanie 3

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.716912	0.719802	0.975513	0.828293
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.885784	0.913961	0.924148	0.918845
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.887745	0.916913	0.923483	0.920041
dataset bez ošetrenia outlierov a škálovania s word frequency	0.869608	0.908208	0.905598	0.906705
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.877941	0.908523	0.918147	0.913119
dataset s ošetrením outlierov	0.734559	0.734187	0.977278	0.837616
dataset s ošetrením outlierov a s ngramami	0.877696	0.918441	0.905865	0.912016
dataset s ošetrením outlierov a s tfidf ngramami	0.856863	0.865267	0.946893	0.903239
dataset s ošetrením outlierov a s word frequency	0.86201	0.879588	0.933977	0.904934
dataset s ošetrením outlierov a s tfidf word frequency	0.850735	0.857915	0.947247	0.899465
dataset so skalovaním	0.722304	0.72548	0.971382	0.830453
dataset so skalovaním a s ngramami	0.87451	0.886704	0.943639	0.913572
dataset so skalovaním a s tfidf ngramami	0.887745	0.90308	0.941568	0.92158

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset so skalovaním a s word frequency	0.863971	0.880074	0.935537	0.906162
dataset so skalovaním a s tfidf word frequency	0.879657	0.89689	0.936197	0.915844
dataset s ošetrením outlierov a skalovaním	0.734559	0.734187	0.977278	0.837616
dataset s ošetrením outlierov a skalovaním a s ngramami	0.87402	0.889184	0.940612	0.913154
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.856863	0.865267	0.946893	0.903239
dataset s ošetrením outlierov a skalovaním a s word frequency	0.86201	0.879588	0.933977	0.904934
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.850735	0.857915	0.947247	0.899465

Tabuľka D.10: Pohlavie - Testovanie 3 SVC

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.744853	0.694932	0.614878	0.613694
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.844118	0.86799	0.756533	0.785509
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.846568	0.863549	0.764501	0.791848
dataset bez ošetrenia outlierov a škálovania s word frequency	0.846569	0.868899	0.761373	0.789835

Varianta datasetu	**accuracy_test**	**precision_test**	**recall_test**	**f1_test**
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.847549	0.877428	0.759728	0.789253
dataset s ošetrením outlierov	0.903431	0.902964	0.862705	0.879256
dataset s ošetrením outlierov a s ngramami	0.853431	0.885027	0.767096	0.798196
dataset s ošetrením outlierov a s tfidf ngramami	0.858333	0.884558	0.778172	0.807098
dataset s ošetrením outlierov a s word frequency	0.87549	0.90335	0.802509	0.832933
dataset s ošetrením outlierov a s tfidf word frequency	0.870588	0.896684	0.796065	0.825883
dataset so skalovaním	0.714461	0.625571	0.561393	0.542525
dataset so skalovaním a s ngramami	0.839215	0.869627	0.745747	0.775642
dataset so skalovaním a s tfidf ngramami	0.843872	0.864394	0.7586	0.786531
dataset so skalovaním a s word frequency	0.847794	0.877909	0.759419	0.789542
dataset so skalovaním a s tfidf word frequency	0.844118	0.871151	0.75563	0.78437
dataset s ošetrením outlierov a skalovaním	0.89951	0.896679	0.859323	0.874798
dataset s ošetrením outlierov a skalovaním a s ngramami	0.858333	0.886317	0.7765	0.806781
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.858578	0.881799	0.779793	0.807957

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a skalovaním a s word frequency	0.875245	0.902242	0.802596	0.832647
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.872549	0.898691	0.798962	0.828795

Tabuľka D.11: Pohlavie - Testovanie 3 RF

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.742157	0.690159	0.630571	0.638602
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.874755	0.869238	0.823024	0.840482
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.870343	0.862341	0.81889	0.835703
dataset bez ošetrenia outlierov a škálovania s word frequency	0.86201	0.851256	0.808701	0.824999
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.868627	0.857732	0.820158	0.834832
dataset s ošetrením outlierov	0.92402	0.919191	0.897458	0.907167
dataset s ošetrením outlierov a s ngramami	0.936029	0.936427	0.909636	0.921511
dataset s ošetrením outlierov a s tfidf ngramami	0.924755	0.926238	0.892578	0.906891
dataset s ošetrením outlierov a s word frequency	0.938235	0.935048	0.916333	0.924909

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a s tfidf word frequency	0.929412	0.92746	0.902707	0.913548
dataset so skalovaním	0.739951	0.68915	0.639543	0.650143
dataset so skalovaním a s ngramami	0.886765	0.878556	0.845481	0.859334
dataset so skalovaním a s tfidf ngramami	0.877941	0.866784	0.835648	0.848591
dataset so skalovaním a s word frequency	0.881618	0.872073	0.839357	0.852897
dataset so skalovaním a s tfidf word frequency	0.878186	0.866191	0.837381	0.849366
dataset s ošetrením outlierov a skalovaním	0.919608	0.913845	0.892056	0.901807
dataset s ošetrením outlierov a skalovaním a s ngramami	0.943382	0.943497	0.920567	0.930888
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.929902	0.930911	0.900542	0.913625
dataset s ošetrením outlierov a skalovaním a s word frequency	0.93897	0.937714	0.915549	0.925531
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.930392	0.930111	0.902407	0.914506

Tabuľka D.12: Pohlavie - Testovanie 3 GBM

D.2. Vek testovanie

Celkovo sme riešili 3 druhy testovania veku.

D.2.1 Testovanie 1

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.326838	0.147735	0.172053	0.138681
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.356434	0.177938	0.194575	0.168154
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.326654	0.147068	0.171753	0.138022
dataset bez ošetrenia outlierov a škálovania s word frequency	0.354596	0.175623	0.19339	0.16646
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.326654	0.147068	0.171753	0.138022
dataset s ošetrením outlierov	0.329228	0.18607	0.174177	0.142593
dataset s ošetrením outlierov a s ngramami	0.360294	0.179803	0.196778	0.170708
dataset s ošetrením outlierov a s tfidf ngramami	0.329044	0.150259	0.173758	0.141471
dataset s ošetrením outlierov a s word frequency	0.359559	0.179144	0.1962	0.16955
dataset s ošetrením outlierov a s tfidf word frequency	0.329044	0.150259	0.173758	0.141471
dataset so skalovaním	0.327022	0.146523	0.171745	0.137478
dataset so skalovaním a s ngramami	0.352206	0.17614	0.192327	0.166068
dataset so skalovaním a s tfidf ngramami	0.327574	0.147284	0.172024	0.137726

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset so skalovaním a s word frequency	0.353125	0.176262	0.192692	0.165907
dataset so skalovaním a s tfidf word frequency	0.327574	0.147284	0.172024	0.137726
dataset s ošetrením outlierov a skalovaním	0.329228	0.18607	0.174177	0.142593
dataset s ošetrením outlierov a skalovaním a s ngramami	0.360294	0.179803	0.196778	0.170708
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.329044	0.150259	0.173758	0.141471
dataset s ošetrením outlierov a skalovaním a s word frequency	0.359559	0.179144	0.1962	0.16955
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.329044	0.150259	0.173758	0.141471

Tabuľka D.13: Vek - Testovanie 1 SVC

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.4	0.351507	0.239796	0.230713
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.397794	0.271267	0.232876	0.219364
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.404044	0.232158	0.233842	0.217882
dataset bez ošetrenia outlierov a škálovania s word frequency	0.406066	0.258321	0.238885	0.225769

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.405883	0.263752	0.237029	0.223061
dataset s ošetrením outlierov	0.709191	0.662523	0.486249	0.500113
dataset s ošetrením outlierov a s ngramami	0.550919	0.401348	0.33192	0.318327
dataset s ošetrením outlierov a s tfidf ngramami	0.545956	0.395001	0.329799	0.316082
dataset s ošetrením outlierov a s word frequency	0.556434	0.379281	0.335378	0.321392
dataset s ošetrením outlierov a s tfidf word frequency	0.548897	0.38088	0.330139	0.316035
dataset so skalovaním	0.397059	0.340925	0.241359	0.234421
dataset so skalovaním a s ngramami	0.39614	0.241702	0.229136	0.21364
dataset so skalovaním a s tfidf ngramami	0.401655	0.24888	0.234928	0.220208
dataset so skalovaním a s word frequency	0.403309	0.250773	0.235129	0.221543
dataset so skalovaním a s tfidf word frequency	0.397978	0.237441	0.232639	0.219104
dataset s ošetrením outlierov a skalovaním	0.712132	0.671613	0.487381	0.500987
dataset s ošetrením outlierov a skalovaním a s ngramami	0.561213	0.373004	0.338819	0.32491
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.555883	0.401994	0.334473	0.320769

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a skalovaním a s word frequency	0.565809	0.433346	0.342434	0.328746
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.545037	0.374371	0.325991	0.310953

Tabuľka D.14: Vek - Testovanie 1 RF

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.406066	0.305323	0.254049	0.251574
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.415809	0.272404	0.255303	0.251064
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.417831	0.269625	0.256214	0.250957
dataset bez ošetrenia outlierov a škálovania s word frequency	0.422427	0.306294	0.264996	0.264974
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.415809	0.278327	0.254947	0.251403
dataset s ošetrením outlierov	0.804228	0.780117	0.637183	0.670974
dataset s ošetrením outlierov a s ngramami	0.752758	0.675379	0.57647	0.601819
dataset s ošetrením outlierov a s tfidf ngramami	0.736765	0.667144	0.55947	0.584646
dataset s ošetrením outlierov a s word frequency	0.745956	0.672844	0.567589	0.594105

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a s tfidf word frequency	0.743015	0.660661	0.564582	0.587986
dataset so skalovaním	0.401838	0.307239	0.254	0.252161
dataset so skalovaním a s ngramami	0.419853	0.279034	0.259132	0.255728
dataset so skalovaním a s tfidf ngramami	0.413235	0.262912	0.251789	0.24661
dataset so skalovaním a s word frequency	0.422978	0.286569	0.261008	0.258925
dataset so skalovaním a s tfidf word frequency	0.421875	0.277683	0.258544	0.254771
dataset s ošetrením outlierov a skalovaním	0.809191	0.783807	0.642613	0.676185
dataset s ošetrením outlierov a skalovaním a s ngramami	0.75239	0.675984	0.575719	0.602053
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.738052	0.656696	0.559835	0.584096
dataset s ošetrením outlierov a skalovaním a s word frequency	0.749816	0.669663	0.569924	0.595567
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.743015	0.651256	0.564449	0.587846

Tabuľka D.15: Vek - Testovanie 1 GBM

D.2.2 Testovanie 2

Pre RFE sme našli tieto najlepšie kombinácie vlastností:

- **SVC dataset bez ošetrenia outlierov a škálovania :** priemerný počet emoji na tweet, priemerný počet URL na tweet, priemerný počet zmienok na tweet, priemerný počet úvodzoviek na tweet, priemerný počet apostrofov na tweet, priemerný počet interpunkčných znamienok na tweet, priemerný počet viacerých interpunkčných znamienok na tweet, priemerná dĺžka slova, priemerný počet viet na tweet, priemerný počet opakovaní na tweet, priemerný počet prísloviek na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet podmetov na tweet, priemerný počet predmetov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet pomocných slovies na tweet, priemerný počet pozitívnych slov na tweet, priemerný počet negatívnych slov na tweet, priemerný počet entít na tweet, priemerný počet negácií na tweet, priemerný počet kognitívnych slov na tweet, priemerný počet zmyslových slov na tweet

- **SVC dataset s ošetrením outlierov :** priemerný počet emoji na tweet, priemerný počet úvodzoviek na tweet, priemerný počet apostrofov na tweet, priemerný počet viacerých interpunkčných znamienok na tweet, priemerná dĺžka slova, priemerný počet viet na tweet, priemerný počet opakovaní na tweet, priemerný počet čiarok na tweet, priemerný počet zámen na tweet, priemerný počet čísloviek na tweet, priemerný počet prísloviek na tweet, priemerný počet predložiek na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet podmetov na tweet, priemerný počet predmetov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet prívlastkov na tweet, priemerný počet pomocných slovies na tweet, priemerný počet negatívnych slov na tweet, priemerný počet negácií na tweet, priemerný počet kognitívnych slov na tweet, priemerný počet zmyslových slov na tweet

- **SVC dataset so skalovaním :** priemerný počet emoji na tweet, priemerný počet úvodzoviek na tweet, priemerný počet viacerých interpunkčných znamienok

na tweet, priemerná dĺžka slova, priemerný počet opakovaní na tweet, percento oznamovacích viet, percento opytovacích viet, percento rozkazovacích viet, percento neštandardných zakončení viet, priemerný počet prídavných mien na tweet, priemerný počet zámen na tweet, priemerný počet častíc na tweet, priemerný počet citosloviec na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet podmetov na tweet, priemerný počet predmetov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet pomocných slovies na tweet, priemerný počet pozitívnych slov na tweet, priemerný počet negatívnych slov na tweet, priemerný počet entít na tweet, priemerný počet negácií na tweet, priemerný počet kognitívnych slov na tweet, priemerný počet zmyslových slov na tweet

- **SVC dataset s ošetrením outlierov a skalovaním :** priemerný počet emoji na tweet, priemerný počet úvodzoviek na tweet, priemerný počet apostrofov na tweet, priemerný počet viacerých interpunkčných znamienok na tweet, priemerná dĺžka slova, priemerný počet viet na tweet, priemerný počet opakovaní na tweet, priemerný počet čiarok na tweet, priemerný počet zámen na tweet, priemerný počet čísloviek na tweet, priemerný počet prísloviek na tweet, priemerný počet predložiek na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet podmetov na tweet, priemerný počet predmetov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet prívlastkov na tweet, priemerný počet pomocných slovies na tweet, priemerný počet negatívnych slov na tweet, priemerný počet negácií na tweet, priemerný počet kognitívnych slov na tweet, priemerný počet zmyslových slov na tweet

- **RF dataset bez ošetrenia outlierov a škálovania :** priemerný počet emoji na tweet, priemerný počet URL na tweet, priemerný počet zmienok na tweet, priemerný počet hashtagov na tweet, priemerný počet číslic na tweet, priemerný počet veľkých písmen na tweet, priemerný počet malých písmen na tweet, priemerný počet apostrofov na tweet, priemerný počet interpunkčných znamienok na tweet, priemerný počet viacerých interpunkčných znamienok na tweet, priemerná dĺžka slova, priemerný počet opakovaní na tweet, veľkosť slovnej zásoby, percento oznamovacích viet, percento opytovacích viet, priemerný počet čiarok

na tweet, priemerný počet zámen na tweet, priemerný počet prísloviek na tweet, priemerný počet častíc na tweet, priemerný počet minulých časov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet prívlastkov na tweet, priemerný počet entít na tweet, priemerný počet negácií na tweet, priemerný počet zmyslových slov na tweet

- **RF dataset s ošetrením outlierov :** priemerný počet emoji na tweet, priemerný počet veľkých písmen na tweet, priemerný počet slov na tweet, priemerný počet znakov na tweet, percento oznamovacích viet, percento opytovacích viet, percento rozkazovacích viet, priemerný počet čiarok na tweet, priemerný počet podstatných mien na tweet, priemerný počet prídavných mien na tweet, priemerný počet slovies na tweet, priemerný počet zámen na tweet, priemerný počet čísloviek na tweet, priemerný počet prísloviek na tweet, priemerný počet predložiek na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet podmetov na tweet, priemerný počet predmetov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet prívlastkov na tweet, priemerný počet pomocných slovies na tweet, priemerný počet negácií na tweet
- **RF dataset so skalovaním :** priemerný počet emoji na tweet, priemerný počet URL na tweet, priemerný počet zmienok na tweet, priemerný počet hashtagov na tweet, priemerný počet číslic na tweet, priemerný počet veľkých písmen na tweet, priemerný počet malých písmen na tweet, priemerný počet apostrofov na tweet, priemerný počet interpunkčných znamienok na tweet, priemerný počet viacerých interpunkčných znamienok na tweet, priemerná dĺžka slova, priemerný počet opakovaní na tweet, veľkosť slovnej zásoby, percento oznamovacích viet, percento opytovacích viet, percento rozkazovacích viet, priemerný počet čiarok na tweet, priemerný počet prísloviek na tweet, priemerný počet častíc na tweet, priemerný počet minulých časov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet neutrálnych slov na tweet, priemerný počet entít na tweet, priemerný počet negácií na tweet, priemerný počet zmyslových slov na tweet

- **RF dataset s ošetrením outlierov a skalovaním :** priemerný počet emoji na tweet, priemerný počet veľkých písmen na tweet, priemerný počet stopslov na tweet, priemerná dĺžka slova, priemerný počet slov na tweet, percento oznamovacích viet, percento opytovacích viet, percento rozkazovacích viet, priemerný počet čiarok na tweet, priemerný počet podstatných mien na tweet, priemerný počet prídavných mien na tweet, priemerný počet slovies na tweet, priemerný počet zámen na tweet, priemerný počet čísloviek na tweet, priemerný počet prísloviek na tweet, priemerný počet predložiek na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet podmetov na tweet, priemerný počet predmetov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet prívlastkov na tweet, priemerný počet pomocných slovies na tweet

- **GBM dataset bez ošetrenia outlierov a škálovania :** priemerný počet emoji na tweet, priemerný počet URL na tweet, priemerný počet zmienok na tweet, priemerný počet hashtagov na tweet, priemerný počet číslic na tweet, priemerný počet veľkých písmen na tweet, priemerný počet apostrofov na tweet, priemerný počet interpunkčných znamienok na tweet, priemerný počet viacerých interpunkčných znamienok na tweet, priemerná dĺžka slova, priemerný počet viet na tweet, priemerný počet opakovaní na tweet, veľkosť slovnej zásoby, percento oznamovacích viet, percento opytovacích viet, percento rozkazovacích viet, priemerný počet podstatných mien na tweet, priemerný počet slovies na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, priemerný počet príslovkových určení na tweet, priemerný počet negatívnych slov na tweet, priemerný počet entít na tweet, priemerný počet negácií na tweet, priemerný počet kognitívnych slov na tweet

- **GBM dataset s ošetrením outlierov :** priemerný počet emoji na tweet, priemerný počet veľkých písmen na tweet, priemerný počet malých písmen na tweet, priemerný počet slov na vetu, priemerný počet slov na tweet, percento oznamovacích viet, percento opytovacích viet, percento rozkazovacích viet, percento iných zakončení viet, priemerný počet čiarok na tweet, priemerný počet podstatných mien na tweet, priemerný počet prídavných mien na tweet, priemerný

počet slovies na tweet, priemerný počet zámen na tweet, priemerný počet číslo-viek na tweet, priemerný počet prísloviek na tweet, priemerný počet predložiek na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, priemerný počet minulých časov na tweet, priemerný počet podmetov na tweet, priemerný počet predmetov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet prívlastkov na tweet, priemerný počet pomocných slovies na tweet

- **GBM dataset so skalovaním :** priemerný počet emoji na tweet, priemerný počet URL na tweet, priemerný počet zmienok na tweet, priemerný počet hashtagov na tweet, priemerný počet číslic na tweet, priemerný počet veľkých písmen na tweet, priemerný počet interpunkčných znamienok na tweet, priemerný počet via-cerých interpunkčných znamienok na tweet, priemerná dĺžka slova, priemerný počet viet na tweet, priemerný počet opakovaní na tweet, veľkosť slovnej zásoby, percento oznamovacích viet, percento opytovacích viet, percento rozkazovacích viet, priemerný počet podstatných mien na tweet, priemerný počet slovies na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, prie-merný počet minulých časov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet pozitívnych slov na tweet, priemerný počet entít na tweet, priemerný počet negácií na tweet, priemerný počet kognitívnych slov na tweet
- **GBM dataset s ošetrením outlierov a skalovaním :** priemerný počet emoji na tweet, priemerný počet veľkých písmen na tweet, priemerný počet slov na vetu, priemerný počet slov na tweet, percento oznamovacích viet, percento opytovacích viet, percento rozkazovacích viet, percento iných zakončení viet, priemerný počet čiarok na tweet, priemerný počet podstatných mien na tweet, priemerný počet prídavných mien na tweet, priemerný počet slovies na tweet, priemerný počet zámen na tweet, priemerný počet číslovieк na tweet, priemerný počet prísloviek na tweet, priemerný počet predložiek na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet podmetov na tweet, priemerný počet predmetov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet prívlastkov na tweet, priemerný počet pomocných slovies na tweet

Varianta datasetu	**accuracy_test**	**precision_test**	**recall_test**	**f1_test**
dataset bez ošetrenia outlierov a škálovania	0.366912	0.190761	0.202632	0.180322
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.415809	0.215303	0.235869	0.214805
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.366544	0.190465	0.199949	0.174177
dataset bez ošetrenia outlierov a škálovania s word frequency	0.408456	0.20932	0.231483	0.210281
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.366544	0.190517	0.199929	0.174159
dataset s ošetrením outlierov	0.461765	0.251049	0.275758	0.258137
dataset s ošetrením outlierov a s ngramami	0.415809	0.215266	0.235869	0.214787
dataset s ošetrením outlierov a s tfidf ngramami	0.467647	0.253954	0.27742	0.258438
dataset s ošetrením outlierov a s word frequency	0.408456	0.20932	0.231483	0.210281
dataset s ošetrením outlierov a s tfidf word frequency	0.468015	0.254354	0.277552	0.258712
dataset so skalovaním	0.359927	0.152024	0.193099	0.162715
dataset so skalovaním a s ngramami	0.415074	0.215014	0.235496	0.214485
dataset so skalovaním a s tfidf ngramami	0.370221	0.157178	0.198835	0.1688

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset so skalovaním a s word frequency	0.408456	0.20932	0.231483	0.210281
dataset so skalovaním a s tfidf word frequency	0.370956	0.156993	0.19897	0.16874
dataset s ošetrením outlierov a skalovaním	0.461765	0.251049	0.275758	0.258137
dataset s ošetrením outlierov a skalovaním a s ngramami	0.415809	0.215266	0.235869	0.214787
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.467647	0.253954	0.27742	0.258438
dataset s ošetrením outlierov a skalovaním a s word frequency	0.408456	0.20932	0.231483	0.210281
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.468015	0.254354	0.277552	0.258712

Tabuľka D.16: Vek - Testovanie 2 RFE SVC

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.401103	0.326152	0.244632	0.238227
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.39853	0.219094	0.231236	0.214192
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.406618	0.232541	0.236431	0.22072
dataset bez ošetrenia outlierov a škálovania s word frequency	0.408824	0.259772	0.238208	0.224179

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.39853	0.229144	0.23157	0.217059
dataset s ošetrením outlierov	0.726471	0.818724	0.516835	0.542117
dataset s ošetrením outlierov a s ngramami	0.565441	0.423159	0.342264	0.327103
dataset s ošetrením outlierov a s tfidf ngramami	0.548162	0.401014	0.330244	0.316288
dataset s ošetrením outlierov a s word frequency	0.555882	0.386083	0.33542	0.320269
dataset s ošetrením outlierov a s tfidf word frequency	0.537868	0.378306	0.321223	0.303461
dataset so skalovaním	0.404044	0.391421	0.24611	0.238514
dataset so skalovaním a s ngramami	0.391177	0.282414	0.226455	0.211567
dataset so skalovaním a s tfidf ngramami	0.403309	0.246685	0.233798	0.217712
dataset so skalovaním a s word frequency	0.396324	0.265726	0.231561	0.217255
dataset so skalovaním a s tfidf word frequency	0.403309	0.232774	0.234678	0.2188
dataset s ošetrením outlierov a skalovaním	0.733824	0.682981	0.521632	0.542209
dataset s ošetrením outlierov a skalovaním a s ngramami	0.553677	0.368722	0.332861	0.315851
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.549265	0.414122	0.331472	0.316464

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a skalovaním a s word frequency	0.557721	0.397856	0.335437	0.320597
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.546691	0.39299	0.327753	0.311321

Tabuľka D.17: Vek - Testovanie 2 RFE RF

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.396324	0.258965	0.247304	0.244107
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.412133	0.271259	0.254314	0.251967
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.413235	0.265941	0.255987	0.251928
dataset bez ošetrenia outlierov a škálovania s word frequency	0.429044	0.298741	0.266564	0.265919
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.414339	0.268335	0.252576	0.249342
dataset s ošetrením outlierov	0.83603	0.809482	0.68832	0.727833
dataset s ošetrením outlierov a s ngramami	0.810662	0.77371	0.6491	0.687246
dataset s ošetrením outlierov a s tfidf ngramami	0.796324	0.734254	0.621778	0.655704
dataset s ošetrením outlierov a s word frequency	0.805883	0.794092	0.636159	0.677961

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a s tfidf word frequency	0.796324	0.733611	0.62648	0.659656
dataset so skalovaním	0.391544	0.254172	0.240529	0.237018
dataset so skalovaním a s ngramami	0.414339	0.28635	0.259654	0.259239
dataset so skalovaním a s tfidf ngramami	0.410294	0.271204	0.255883	0.252748
dataset so skalovaním a s word frequency	0.420956	0.299625	0.264503	0.265513
dataset so skalovaním a s tfidf word frequency	0.415074	0.280161	0.258508	0.257105
dataset s ošetrením outlierov a skalovaním	0.8375	0.808254	0.683924	0.7228
dataset s ošetrením outlierov a skalovaním a s ngramami	0.814706	0.775631	0.654152	0.692087
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.799633	0.744789	0.627019	0.661733
dataset s ošetrením outlierov a skalovaním a s word frequency	0.797795	0.74536	0.6256	0.660807
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.804044	0.751232	0.631074	0.666908

Tabuľka D.18: Vek - Testovanie 2 RFE GBM

Najlepšie kombinácie vlastností pre FS metódu:

- **SVC dataset bez ošetrenia outlierov a škálovania :** priemerný počet nealfabetických znakov na tweet, priemerný počet znakov na tweet, priemerný počet

veľkých písmen na

- **SVC dataset s ošetrením outlierov :** priemerný počet priamych predmetov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet prívlastkov na tweet, priemerný počet pomocných slovies na tweet
- **SVC dataset so skalovaním :** priemerný počet nealfabetických znakov na tweet, priemerný počet znakov na tweet, priemerný počet veľkých písmen na tweet, priemerný počet slov na vetu
- **SVC dataset s ošetrením outlierov a skalovaním :** priemerný počet priamych predmetov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet menných prívlastkov na tweet, priemerný počet pomocných slovies na tweet
- **RF dataset bez ošetrenia outlierov a škálovania :** priemerný počet slov na tweet, priemerný počet interpunkčných znamienok na tweet, priemerný počet emoji na tweet, priemerný počet opakovaní na tweet, priemerný počet zmienok na tweet, priemerný počet veľkých písmen na tweet, priemerný počet úvodzoviek na tweet, priemerný počet kognitívnych slov na tweet, priemerný počet slov na vetu, priemerný počet podmetov na tweet
- **RF dataset s ošetrením outlierov :** priemerný počet priamych predmetov na tweet, priemerný počet častíc na tweet, percento rozkazovacích viet, priemerný počet podstatných mien na tweet, priemerný počet prísloviek na tweet, priemerný počet príslovkových určení na tweet, priemerný počet podmetov na tweet, priemerný počet znakov na tweet
- **RF dataset so skalovaním :** priemerný počet slov na tweet, priemerný počet interpunkčných znamienok na tweet, percento oznamovacích viet, priemerný počet opakovaní na tweet, priemerná dĺžka slova, priemerný počet úvodzoviek na tweet, priemerný počet príslovkových určení na tweet
- **RF dataset s ošetrením outlierov a skalovaním :** priemerný počet priamych predmetov na tweet, priemerný počet častíc na tweet, percento rozkazovacích viet, priemerný počet čiarok na tweet, priemerný počet slovies na tweet, priemerný počet podstatných mien na tweet

- **GBM dataset bez ošetrenia outlierov a škálovania :** priemerný počet malých písmen na tweet, priemerný počet negácií na tweet, priemerný počet viet na tweet, priemerný počet emoji na tweet, priemerný počet častíc na tweet, priemerný počet negatívnych slov na tweet, percento opytovacích viet, priemerný počet nealfabetických znakov na tweet, priemerný počet podmetov na tweet, priemerný počet zmyslových slov na tweet, priemerný počet menných prívlastkov na tweet
- **GBM dataset s ošetrením outlierov :** priemerný počet priamych predmetov na tweet, priemerný počet častíc na tweet, percento oznamovacích viet, percento opytovacích viet, priemerný počet čísloviek na tweet, priemerný počet čiarok na tweet, priemerný počet príslovkových určení na tweet, percento iných zakončení viet, priemerný počet zámen na tweet, priemerný počet prídavných mien na tweet, priemerný počet spojok na tweet
- **GBM dataset so skalovaním :** priemerný počet malých písmen na tweet, priemerný počet negácií na tweet, priemerný počet viet na tweet, priemerný počet prídavných mien na tweet, priemerný počet opakovaní na tweet, priemerný počet emoji na tweet, priemerná dĺžka slova, priemerný počet menných prívlastkov na tweet, priemerný počet interpunkčných znamienok na tweet, percento iných zakončení viet
- **GBM dataset s ošetrením outlierov a skalovaním :** priemerný počet priamych predmetov na tweet, priemerný počet častíc na tweet, percento oznamovacích viet, percento opytovacích viet, priemerný počet čísloviek na tweet, priemerný počet čiarok na tweet, priemerný počet príslovkových určení na tweet, percento iných zakončení viet, priemerný počet zámen na tweet, priemerný počet prídavných mien na tweet, priemerný počet spojok na tweet

Varianta datasetu	**accuracy_test**	**precision_test**	**recall_test**	**f1_test**
dataset bez ošetrenia outlierov a škálovania	0.37353	0.170276	0.199738	0.167534
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.404044	0.272465	0.228286	0.208354

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.369485	0.169289	0.196125	0.164122
dataset bez ošetrenia outlierov a škálovania s word frequency	0.403309	0.274634	0.228592	0.20931
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.369485	0.168838	0.196119	0.164143
dataset s ošetrením outlierov	0.44853	0.300316	0.264585	0.247394
dataset s ošetrením outlierov a s ngramami	0.411765	0.277392	0.232661	0.213466
dataset s ošetrením outlierov a s tfidf ngramami	0.398162	0.192189	0.221794	0.198143
dataset s ošetrením outlierov a s word frequency	0.409927	0.277816	0.231978	0.212967
dataset s ošetrením outlierov a s tfidf word frequency	0.399265	0.193627	0.222539	0.198914
dataset so skalovaním	0.377206	0.194966	0.203116	0.173845
dataset so skalovaním a s ngramami	0.403309	0.271849	0.227812	0.207795
dataset so skalovaním a s tfidf ngramami	0.369486	0.212427	0.197182	0.16753
dataset so skalovaním a s word frequency	0.404412	0.274683	0.229051	0.209546
dataset so skalovaním a s tfidf word frequency	0.369118	0.212034	0.196919	0.167215
dataset s ošetrením outlierov a skalovaním	0.44853	0.300316	0.264585	0.247394

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a skalovaním a s ngramami	0.411765	0.277392	0.232661	0.213466
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.398162	0.192189	0.221794	0.198143
dataset s ošetrením outlierov a skalovaním a s word frequency	0.409927	0.277816	0.231978	0.212967
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.399265	0.193627	0.222539	0.198914

Tabuľka D.19: Vek - Testovanie 2 FS SVC

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.38603	0.343924	0.237908	0.232457
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.403677	0.295285	0.235005	0.220424
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.405883	0.259943	0.240542	0.228474
dataset bez ošetrenia outlierov a škálovania s word frequency	0.405515	0.266046	0.236942	0.22534
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.405515	0.299087	0.236408	0.222453
dataset s ošetrením outlierov	0.726103	0.699875	0.536434	0.562902
dataset s ošetrením outlierov a s ngramami	0.476471	0.316102	0.277635	0.263227

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a s tfidf ngramami	0.470956	0.32562	0.277579	0.264085
dataset s ošetrením outlierov a s word frequency	0.472794	0.313371	0.274322	0.258049
dataset s ošetrením outlierov a s tfidf word frequency	0.468015	0.309539	0.277988	0.266137
dataset so skalovaním	0.367647	0.236746	0.222859	0.215762
dataset so skalovaním a s ngramami	0.401471	0.24287	0.231767	0.217081
dataset so skalovaním a s tfidf ngramami	0.395956	0.259502	0.231827	0.219188
dataset so skalovaním a s word frequency	0.404044	0.27468	0.237242	0.227267
dataset so skalovaním a s tfidf word frequency	0.404412	0.272285	0.239529	0.228526
dataset s ošetrením outlierov a skalovaním	0.705882	0.784513	0.516689	0.541625
dataset s ošetrením outlierov a skalovaním a s ngramami	0.492647	0.331577	0.28975	0.273903
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.481618	0.327118	0.28665	0.275626
dataset s ošetrením outlierov a skalovaním a s word frequency	0.476471	0.278712	0.280148	0.265124
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.470588	0.304645	0.278042	0.264981

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test

Tabuľka D.20: Vek - Testovanie 2 FS RF

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.41728	0.330082	0.273738	0.276866
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.426838	0.308045	0.271661	0.272129
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.420956	0.287566	0.266793	0.268041
dataset bez ošetrenia outlierov a škálovania s word frequency	0.415809	0.261471	0.251466	0.246657
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.423162	0.281266	0.266166	0.264043
dataset s ošetrením outlierov	0.822574	0.803025	0.6306307	0.710767
dataset s ošetrením outlierov a s ngramami	0.790441	0.730066	0.619107	0.65158
dataset s ošetrením outlierov a s tfidf ngramami	0.779412	0.774925	0.602792	0.640402
dataset s ošetrením outlierov a s word frequency	0.784927	0.729848	0.604963	0.638261
dataset s ošetrením outlierov a s tfidf word frequency	0.769853	0.703091	0.594938	0.622102
dataset so skalovaním	0.403677	0.307926	0.262878	0.263713
dataset so skalovaním a s ngramami	0.422059	0.306504	0.266984	0.268659

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset so skalovaním a s tfidf ngramami	0.424633	0.281239	0.26494	0.264239
dataset so skalovaním a s word frequency	0.433088	0.290708	0.268002	0.267374
dataset so skalovaním a s tfidf word frequency	0.419485	0.27209	0.2613	0.259082
dataset s ošetrením outlierov a skalovaním	0.829559	0.802573	0.631176	0.724542
dataset s ošetrením outlierov a skalovaním a s ngramami	0.799633	0.739199	0.629443	0.663263
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.78603	0.739322	0.611494	0.646084
dataset s ošetrením outlierov a skalovaním a s word frequency	0.790441	0.749331	0.620339	0.656649
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.777941	0.709446	0.607139	0.637538

Tabuľka D.21: Vek - Testovanie 2 FS GBM

D.2.3 Testovanie 3

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.426075	0.222206	0.245791	0.224282
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.419956	0.307658	0.31822	0.318849

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.376146	0.175726	0.208714	0.175255
dataset bez ošetrenia outlierov a škálovania s word frequency	0.366862	0.269483	0.280965	0.273059
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.381993	0.179477	0.207606	0.176987
dataset s ošetrením outlierov	0.504693	0.342122	0.314657	0.311145
dataset s ošetrením outlierov a s ngramami	0.34994	0.118006	0.173365	0.128124
dataset s ošetrením outlierov a s tfidf ngramami	0.491177	0.255035	0.297592	0.278466
dataset s ošetrením outlierov a s word frequency	0.302787	0.041999	0.143147	0.06663
dataset s ošetrením outlierov a s tfidf word frequency	0.483027	0.259077	0.302035	0.28221
dataset so skalovaním	0.399792	0.201962	0.224617	0.209669
dataset so skalovaním a s ngramami	0.424552	0.323063	0.31748	0.312518
dataset so skalovaním a s tfidf ngramami	0.425424	0.325772	0.26739	0.266457
dataset so skalovaním a s word frequency	0.383008	0.278492	0.281975	0.28536
dataset so skalovaním a s tfidf word frequency	0.439254	0.337048	0.27396	0.28123
dataset s ošetrením outlierov a skalovaním	0.507452	0.335869	0.322045	0.311166

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a skalovaním a s ngramami	0.337149	0.117518	0.175525	0.127707
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.490107	0.260203	0.303585	0.280551
dataset s ošetrením outlierov a skalovaním a s word frequency	0.298979	0.042642	0.140776	0.064989
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.498	0.268557	0.305716	0.281402

Tabuľka D.22: Vek - Testovanie 3 SVC

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.427044	0.2519	0.239386	0.2316
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.423606	0.202924	0.229256	0.21617
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.415237	0.200145	0.229462	0.214728
dataset bez ošetrenia outlierov a škálovania s word frequency	0.406456	0.211591	0.234084	0.218824
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.416986	0.287458	0.246348	0.224127
dataset s ošetrením outlierov	0.737139	0.698482	0.53422	0.552605
dataset s ošetrením outlierov a s ngramami	0.521432	0.279225	0.30192	0.294277

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset s ošetrením outlierov a s tfidf ngramami	0.594274	0.467886	0.355978	0.346014
dataset s ošetrením outlierov a s word frequency	0.63172	0.391876	0.36578	0.355568
dataset s ošetrením outlierov a s tfidf word frequency	0.56882	0.414222	0.345049	0.336651
dataset so skalovaním	0.410268	0.372409	0.244042	0.2313
dataset so skalovaním a s ngramami	0.416601	0.349411	0.242487	0.222699
dataset so skalovaním a s tfidf ngramami	0.417508	0.20164	0.236733	0.214516
dataset so skalovaním a s word frequency	0.418234	0.206479	0.23824	0.212956
dataset so skalovaním a s tfidf word frequency	0.423449	0.20686	0.241317	0.220498
dataset s ošetrením outlierov a skalovaním	0.757394	0.688648	0.551352	0.553957
dataset s ošetrením outlierov a skalovaním a s ngramami	0.539765	0.268229	0.307988	0.28111
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.574237	0.457799	0.353459	0.33011
dataset s ošetrením outlierov a skalovaním a s word frequency	0.615261	0.487929	0.380814	0.367074
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.600296	0.46181	0.361174	0.331273

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test

Tabuľka D.23: Vek - Testovanie 3 RF

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset bez ošetrenia outlierov a škálovania	0.404998	0.268762	0.240432	0.237384
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.424813	0.396173	0.249282	0.244863
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.393493	0.271929	0.235052	0.229137
dataset bez ošetrenia outlierov a škálovania s word frequency	0.428613	0.317153	0.255814	0.25697
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.410418	0.273885	0.246308	0.240322
dataset s ošetrením outlierov	0.836027	0.806036	0.668409	0.704849
dataset s ošetrením outlierov a s ngramami	0.805321	0.756951	0.621336	0.685593
dataset s ošetrením outlierov a s tfidf ngramami	0.824759	0.793737	0.640162	0.675395
dataset s ošetrením outlierov a s word frequency	0.833134	0.782417	0.646552	0.693907
dataset s ošetrením outlierov a s tfidf word frequency	0.787608	0.771625	0.643621	0.661407
dataset so skalovaním	0.392419	0.238006	0.237289	0.225544
dataset so skalovaním a s ngramami	0.429216	0.285182	0.264252	0.256933

Varianta datasetu	accuracy_test	precision_test	recall_test	f1_test
dataset so skalovaním a s tfidf ngramami	0.413723	0.330086	0.257662	0.253592
dataset so skalovaním a s word frequency	0.42889	0.2954	0.259143	0.250096
dataset so skalovaním a s tfidf word frequency	0.408355	0.234974	0.241529	0.234222
dataset s ošetrením outlierov a skalovaním	0.828871	0.807739	0.675656	0.699144
dataset s ošetrením outlierov a skalovaním a s ngramami	0.812466	0.795631	0.625871	0.668919
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.82686	0.798031	0.639279	0.673174
dataset s ošetrením outlierov a skalovaním a s word frequency	0.821949	0.819653	0.663766	0.713084
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.809592	0.79795	0.64694	0.672938

Tabuľka D.24: Vek - Testovanie 3 GBM

D.3. Hypotezy

Výsledné tabuľky testovania za účelom potvrdenia alebo zamietnutia hypotéz.

D.3.1 Testovanie iba so syntaktickými črtami

Varianta datasetu	**accuracy_test**
bez ošetrenia outlierov a škálovania	0.345588
bez ošetrenia outlierov a škálovania s ngramami	0.409559
bez ošetrenia outlierov a škálovania s tfidf ngramami	0.347059
s ošetrením outlierov	0.396324
s ošetrením outlierov a s ngramami	0.413971
s ošetrením outlierov a s tfidf ngramami	0.397059
so skalovaním	0.369853
so skalovaním a s ngramami	0.401471
so skalovaním a s tfidf ngramami	0.411029
s ošetrením outlierov a skalovaním	0.396324
s ošetrením outlierov a skalovaním a s ngramami	0.413971
s ošetrením outlierov a skalovaním a s tfidf ngramami	0.397059

Tabuľka D.25: Syntaktické - Testovanie 1 SVC

Varianta datasetu	**accuracy_test**
bez ošetrenia outlierov a škálovania	0.346324
bez ošetrenia outlierov a škálovania s ngramami	0.405882
bez ošetrenia outlierov a škálovania s tfidf ngramami	0.383824
s ošetrením outlierov	0.732353
s ošetrením outlierov a s ngramami	0.610294
s ošetrením outlierov a s tfidf ngramami	0.572794

Varianta datasetu	**accuracy_test**
so skalovaním	0.355882
so skalovaním a s ngramami	0.408088
so skalovaním a s tfidf ngramami	0.406618
s ošetrením outlierov a skalovaním	0.736029
s ošetrením outlierov a skalovaním a s ngramami	0.614706
s ošetrením outlierov a skalovaním a s tfidf ngramami	0.595588

Tabuľka D.26: Syntaktické - Testovanie 1 RF

Varianta datasetu	**accuracy_test**
bez ošetrenia outlierov a škálovania	0.360662
bez ošetrenia outlierov a škálovania s ngramami	0.407353
bez ošetrenia outlierov a škálovania s tfidf ngramami	0.390442
s ošetrením outlierov	0.797059
s ošetrením outlierov a s ngramami	0.726103
s ošetrením outlierov a s tfidf ngramami	0.705515
so skalovaním	0.367279
so skalovaním a s ngramami	0.410662
so skalovaním a s tfidf ngramami	0.406618
s ošetrením outlierov a skalovaním	0.798162
s ošetrením outlierov a skalovaním a s ngramami	0.728309
s ošetrením outlierov a skalovaním a s tfidf ngramami	0.71728

Varianta datasetu	**accuracy_test**

Tabuľka D.27: Syntaktické - Testovanie 1 GBM

Varianta	**accuracy_test**
bez ošetrenia outlierov a škálovania	0.370588
bez ošetrenia outlierov a škálovania s ngramami	0.405148
bez ošetrenia outlierov a škálovania s tfidf ngramami	0.394118
s ošetrením outlierov	0.66875
s ošetrením outlierov a s ngramami	0.621692
s ošetrením outlierov a s tfidf ngramami	0.655515
so skalovaním	0.373897
so skalovaním a s ngramami	0.407353
so skalovaním a s tfidf ngramami	0.409192
s ošetrením outlierov a skalovaním	0.668015
s ošetrením outlierov a skalovaním a s ngramami	0.621692
s ošetrením outlierov a skalovaním a s tfidf ngramami	0.655883

Tabuľka D.28: Syntaktické - Testovanie 2 SVC

Varianta datasetu	**accuracy_test**
bez ošetrenia outlierov a škálovania	0.344118
bez ošetrenia outlierov a škálovania s ngramami	0.413235

Varianta datasetu	**accuracy_test**
bez ošetrenia outlierov a škálovania s tfidf ngramami	0.416176
s ošetrením outlierov	0.711765
s ošetrením outlierov a s ngramami	0.580147
s ošetrením outlierov a s tfidf ngramami	0.556618
so skalovaním	0.352941
so skalovaním a s ngramami	0.402941
so skalovaním a s tfidf ngramami	0.419118
s ošetrením outlierov a skalovaním	0.717647
s ošetrením outlierov a skalovaním a s ngramami	0.571324
s ošetrením outlierov a skalovaním a s tfidf ngramami	0.547059

Tabuľka D.29: Syntaktické - Testovanie 2 RF

Varianta	**accuracy_test**
bez ošetrenia outlierov a škálovania	0.347059
bez ošetrenia outlierov a škálovania s ngramami	0.409559
bez ošetrenia outlierov a škálovania s tfidf ngramami	0.400735
s ošetrením outlierov	0.834559
s ošetrením outlierov a s ngramami	0.811765
s ošetrením outlierov a s tfidf ngramami	0.802206
so skalovaním	0.347794
so skalovaním a s ngramami	0.413235

Varianta	**accuracy_test**
so skalovaním a s tfidf ngramami	0.401471
s ošetrením outlierov a skalovaním	0.833824
s ošetrením outlierov a skalovaním a s ngramami	0.809559
s ošetrením outlierov a skalovaním a s tfidf ngramami	0.804412

Tabuľka D.30: Syntaktické - Testovanie 2 GBM

Varianta datasetu	**accuracy_test**
bez ošetrenia outlierov a škálovania	0.376471
bez ošetrenia outlierov a škálovania s ngramami	0.395588
bez ošetrenia outlierov a škálovania s tfidf ngramami	0.427206
s ošetrením outlierov	0.4875
s ošetrením outlierov a s ngramami	0.384559
s ošetrením outlierov a s tfidf ngramami	0.497794
so skalovaním	0.377941
so skalovaním a s ngramami	0.399265
so skalovaním a s tfidf ngramami	0.402941
s ošetrením outlierov a skalovaním	0.4875
s ošetrením outlierov a skalovaním a s ngramami	0.384559
s ošetrením outlierov a skalovaním a s tfidf ngramami	0.497794

Tabuľka D.31: Syntaktické - Testovanie 3 SVC

Varianta datasetu	**accuracy_test**
bez ošetrenia outlierov a škálovania	0.364339
bez ošetrenia outlierov a škálovania s ngramami	0.407353
bez ošetrenia outlierov a škálovania s tfidf ngramami	0.420956
s ošetrením outlierov	0.604412
s ošetrením outlierov a s ngramami	0.493383
s ošetrením outlierov a s tfidf ngramami	0.534927
so skalovaním	0.373162
so skalovaním a s ngramami	0.399633
so skalovaním a s tfidf ngramami	0.399265
s ošetrením outlierov a skalovaním	0.606618
s ošetrením outlierov a skalovaním a s ngramami	0.4875
s ošetrením outlierov a skalovaním a s tfidf ngramami	0.537868

Tabuľka D.32: Syntaktické - Testovanie 3 RF

Varianta datasetu	**accuracy_test**
bez ošetrenia outlierov a škálovania	0.353676
bez ošetrenia outlierov a škálovania s ngramami	0.405882
bez ošetrenia outlierov a škálovania s tfidf ngramami	0.394853
s ošetrením outlierov	0.811029
s ošetrením outlierov a s ngramami	0.792647
s ošetrením outlierov a s tfidf ngramami	0.784559

Varianta datasetu	accuracy_test
so skalovaním	0.357353
so skalovaním a s ngramami	0.411765
so skalovaním a s tfidf ngramami	0.390441
s ošetrením outlierov a skalovaním	0.816176
s ošetrením outlierov a skalovaním a s ngramami	0.783824
s ošetrením outlierov a skalovaním a s tfidf ngramami	0.755147

Tabuľka D.33: Syntaktické - Testovanie 3 GBM

D.3.2 Testovanie iba so lexikálnymi črtami

Varianta datasetu	accuracy_test
bez ošetrenia outlierov a škálovania	0.32978
bez ošetrenia outlierov a škálovania s frekvenciou slov	0.334927
bez ošetrenia outlierov a škálovania s tfidf frekvenciou slov	0.32978
s ošetrením outlierov	0.416544
s ošetrením outlierov a s frekvenciou slov	0.422794
s ošetrením outlierov a s tfidf frekvenciou slov	0.420221
so skalovaním	0.397795
so skalovaním a s frekvenciou slov	0.433824
so skalovaním a s tfidf frekvenciou slov	0.450736
s ošetrením outlierov a skalovaním	0.416544

Varianta datasetu	**accuracy_test**
s ošetrením outlierov a skalovaním a s frekvenciou slov	0.422794
s ošetrením outlierov a skalovaním a s tfidf frekvenciou slov	0.420221

Tabuľka D.34: Lexikálne - Testovanie 1 SVC

Varianta datasetu	**accuracy_test**
bez ošetrenia outlierov a škálovania	0.405515
bez ošetrenia outlierov a škálovania s frekvenciou slov	0.407353
bez ošetrenia outlierov a škálovania s tfidf frekvenciou slov	0.411397
s ošetrením outlierov	0.758824
s ošetrením outlierov a s frekvenciou slov	0.636765
s ošetrením outlierov a s tfidf frekvenciou slov	0.621324
so skalovaním	0.403677
so skalovaním a s frekvenciou slov	0.408456
so skalovaním a s tfidf frekvenciou slov	0.402941
s ošetrením outlierov a skalovaním	0.75625
s ošetrením outlierov a skalovaním a s frekvenciou slov	0.632721
s ošetrením outlierov a skalovaním a s tfidf frekvenciou slov	0.615442

Tabuľka D.35: Lexikálne - Testovanie 1 RF

Varianta datasetu	**accuracy_test**
bez ošetrenia outlierov a škálovania	0.393383
bez ošetrenia outlierov a škálovania s frekvenciou slov	0.419853
bez ošetrenia outlierov a škálovania s tfidf frekvenciou slov	0.42353
s ošetrením outlierov	0.878677
s ošetrením outlierov a s frekvenciou slov	0.857353
s ošetrením outlierov a s tfidf frekvenciou slov	0.849633
so skalovaním	0.392647
so skalovaním a s frekvenciou slov	0.424265
so skalovaním a s tfidf frekvenciou slov	0.416912
s ošetrením outlierov a skalovaním	0.877942
s ošetrením outlierov a skalovaním a s frekvenciou slov	0.859192
s ošetrením outlierov a skalovaním a s tfidf frekvenciou slov	0.849632

Tabuľka D.36: Lexikálne - Testovanie 1 GBM

Varianta datasetu	**accuracy_test**
bez ošetrenia outlierov a škálovania	0.372427
bez ošetrenia outlierov a škálovania s frekvenciou slov	0.434559
bez ošetrenia outlierov a škálovania s tfidf frekvenciou slov	0.381985
s ošetrením outlierov	0.491544
s ošetrením outlierov a s frekvenciou slov	0.434191
s ošetrením outlierov a s tfidf frekvenciou slov	0.491912

Varianta datasetu	**accuracy_test**
so skalovaním	0.395956
so skalovaním a s frekvenciou slov	0.433456
so skalovaním a s tfidf frekvenciou slov	0.450736
s ošetrením outlierov a skalovaním	0.491544
s ošetrením outlierov a skalovaním a s frekvenciou slov	0.434191
s ošetrením outlierov a skalovaním a s tfidf frekvenciou slov	0.491912

Tabuľka D.37: Lexikálne - Testovanie 2 SVC

Varianta datasetu	**accuracy_test**
bez ošetrenia outlierov a škálovania	0.390074
bez ošetrenia outlierov a škálovania s frekvenciou slov	0.413603
bez ošetrenia outlierov a škálovania s tfidf frekvenciou slov	0.410294
s ošetrením outlierov	0.73125
s ošetrením outlierov a s frekvenciou slov	0.577574
s ošetrením outlierov a s tfidf frekvenciou slov	0.574265
so skalovaním	0.395221
so skalovaním a s frekvenciou slov	0.405147
so skalovaním a s tfidf frekvenciou slov	0.40625
s ošetrením outlierov a skalovaním	0.737868
s ošetrením outlierov a skalovaním a s frekvenciou slov	0.575736
s ošetrením outlierov a skalovaním a s tfidf frekvenciou slov	0.566912

Varianta datasetu	accuracy_test

Tabuľka D.38: Lexikálne - Testovanie 2 RF

Varianta datasetu	accuracy_test
bez ošetrenia outlierov a škálovania	0.386397
bez ošetrenia outlierov a škálovania s frekvenciou slov	0.421692
bez ošetrenia outlierov a škálovania s tfidf frekvenciou slov	0.415074
s ošetrením outlierov	0.841177
s ošetrením outlierov a s frekvenciou slov	0.819853
s ošetrením outlierov a s tfidf frekvenciou slov	0.813236
so skalovaním	0.388236
so skalovaním a s frekvenciou slov	0.428309
so skalovaním a s tfidf frekvenciou slov	0.416545
s ošetrením outlierov a skalovaním	0.844485
s ošetrením outlierov a skalovaním a s frekvenciou slov	0.819118
s ošetrením outlierov a skalovaním a s tfidf frekvenciou slov	0.813971

Tabuľka D.39: Lexikálne - Testovanie 2 GBM

Varianta datasetu	accuracy_test
bez ošetrenia outlierov a škálovania	0.404044
bez ošetrenia outlierov a škálovania s frekvenciou slov	0.383089

Varianta datasetu	**accuracy_test**
bez ošetrenia outlierov a škálovania s tfidf frekvenciou slov	0.388236
s ošetrením outlierov	0.51728
s ošetrením outlierov a s frekvenciou slov	0.382721
s ošetrením outlierov a s tfidf frekvenciou slov	0.436765
so skalovaním	0.401471
so skalovaním a s frekvenciou slov	0.41875
so skalovaním a s tfidf frekvenciou slov	0.443015
s ošetrením outlierov a skalovaním	0.51728
s ošetrením outlierov a skalovaním a s frekvenciou slov	0.382721
s ošetrením outlierov a skalovaním a s tfidf frekvenciou slov	0.436765

Tabuľka D.40: Lexikálne - Testovanie 3 SVC

Varianta datasetu	**accuracy_test**
bez ošetrenia outlierov a škálovania	0.407353
bez ošetrenia outlierov a škálovania s frekvenciou slov	0.406618
bez ošetrenia outlierov a škálovania s tfidf frekvenciou slov	0.401471
s ošetrením outlierov	0.72978
s ošetrením outlierov a s frekvenciou slov	0.598162
s ošetrením outlierov a s tfidf frekvenciou slov	0.567647
so skalovaním	0.408824
so skalovaním a s frekvenciou slov	0.415441

Varianta datasetu	**accuracy_test**
so skalovaním a s tfidf frekvenciou slov	0.415809
s ošetrením outlierov a skalovaním	0.740442
s ošetrením outlierov a skalovaním a s frekvenciou slov	0.597794
s ošetrením outlierov a skalovaním a s tfidf frekvenciou slov	0.582353

Tabuľka D.41: Lexikálne - Testovanie 3 RF

Varianta datasetu	**accuracy_test_avg**
bez ošetrenia outlierov a škálovania	0.382631
bez ošetrenia outlierov a škálovania s frekvenciou slov	0.411809
bez ošetrenia outlierov a škálovania s tfidf frekvenciou slov	0.415003
s ošetrením outlierov	0.83744
s ošetrením outlierov a s frekvenciou slov	0.8128
s ošetrením outlierov a s tfidf frekvenciou slov	0.804711
so skalovaním	0.388043
so skalovaním a s frekvenciou slov	0.407513
so skalovaním a s tfidf frekvenciou slov	0.392506
s ošetrením outlierov a skalovaním	0.835801
s ošetrením outlierov a skalovaním a s frekvenciou slov	0.807242
s ošetrením outlierov a skalovaním a s tfidf frekvenciou slov	0.807965

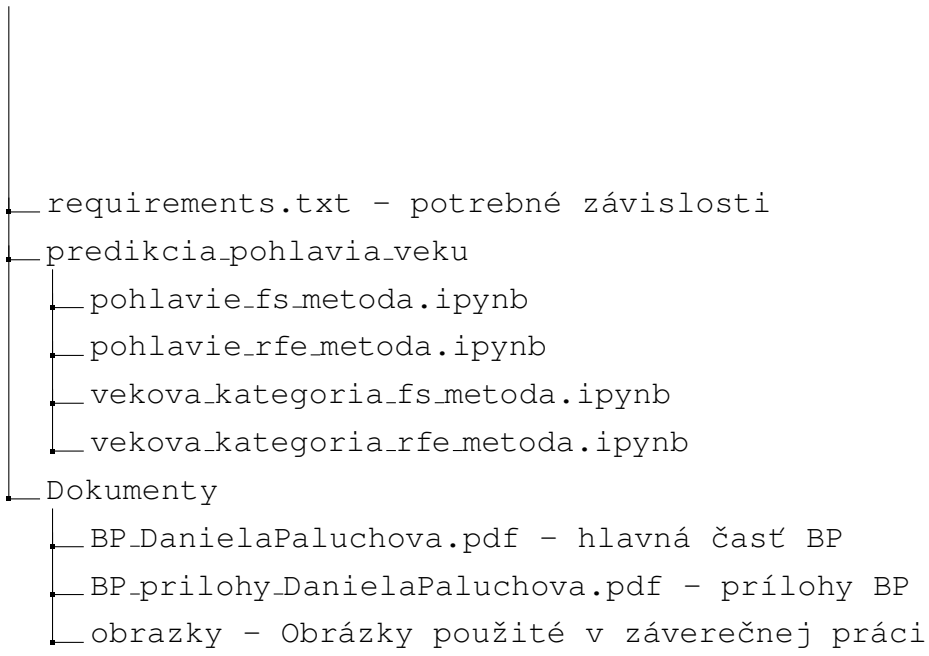
Tabuľka D.42: Lexikálne - Testovanie 3 GBM

Príloha E

Opis digitálnej časti práce

Evidenčné číslo práce v evidenčnom systéme: FIIT-100241-110868
Obsah digitálnej časti práce (archív ZIP):

```
BP_DanielaPaluchova.zip
├── datasety
│   ├── dataset_ocisteny.json
│   ├── lexikalne_crty_ocisteny.json
│   ├── semanticke_crty_ocisteny.json
│   └── syntakticke_crty_ocisteny.json
├── extrakcia_crt - Extrakcia čŕt z očisteného datasetu
│   ├── syntakticke_crty_extrakcia.ipynb
│   ├── semanticke_crty_extrakcia.ipynb
│   └── lexikalne_crty_extrakcia.ipynb
├── analyza_cistenie_datasetu
│   ├── analyza_neocisteneho.ipynb
│   ├── analyza_ocisteneho.ipynb
│   └── cistenie_datasetu.ipynb
└── hypotezy
    ├── hypoteza1_2_3.ipynb
    ├── hypoteza5_lexikalne_crty.ipynb
    └── hypoteza5_syntakticke_crty.ipynb
```


requirements.txt – potrebné závislosti
predikcia_pohlavia_veku
pohlavie_fs_metoda.ipynb
pohlavie_rfe_metoda.ipynb
vekova_kategoria_fs_metoda.ipynb
vekova_kategoria_rfe_metoda.ipynb
Dokumenty
BP_DanielaPaluchova.pdf – hlavná časť BP
BP_prilohy_DanielaPaluchova.pdf – prílohy BP
obrazky – Obrázky použité v záverečnej práci