D22 - Algoritmus na nájdenie odľahlých hodnôt

Projekt financovaný Európskou úniou Next GenerationEU prostredníctvom Plánu obnovy a odolnosti SR v rámci projektu č. 09I05-03-V02-00049.



ÚVOD

Projekt „**Automatizácia analýzy právnych textov na základe strojového učenia**“ (ďalej len „**ALTAML**“) predstavuje kľúčovú iniciatívu zameranú na integráciu inovatívnych prístupov v oblasti spracovania údajov a následnej analýzy, so zameraní na oblasť práva, ktorá zahŕňa aj právo informačných a komunikačných technológií. Cieľom tohto projektu je vyvinúť a overiť účinné metódy automatizovanej analýzy právnych textov pomocou techník strojového učenia. To zahŕňa najmä vývoj nástrojov, ktoré uľahčujú spracovanie a analýzu veľkých objemov údajov vo forme rôznych právnych dokumentov, ako aj extrakciu relevantných informácií (atribútov) z takýchto dokumentov vrátane identifikácie kľúčových pojmov, odkazov na právne predpisy a ďalších atribútov.

Cieľom projektu ALTAML je teda prispieť k efektívnejšiemu prístupu k právnym informáciám a k zrýchleniu právnych procesov, čím sa zabezpečí vyššia miera právnej istoty a zlepší sa dostupnosť právnych textov, výsledkov ich analýzy a relevantných právnych informácií pre odbornú aj širokú verejnosť.

V rámci pracovného balíka KPB4 bola vyvinutá **D22 – Algoritmus na nájdenie odľahlých hodnôt**. Tento dokument prezentuje navrhnutý prístup k identifikácii takýchto textov pomocou metód strojového učenia, konkrétne s využitím modelu Isolation Forest založeného na učení bez učiteľa a následnej interpretácie výsledkov pomocou metódy SHAP.

V kontexte tejto úlohy sú netypické právne texty identifikované prostredníctvom detekcie tzv. outlierov (odľahlých prvkov dátovej sady). Outlier predstavuje pozorovanie v dátach, ktoré sa výrazne odlišuje od ostatných a nezodpovedá bežnému vzoru správania v danej dátovej sade. Takéto prípady môžu indikovať zriedkavé, neštandardné alebo inak významné právne rozhodnutia.

Využitím metódy SHAP je možné vysvetliť, prečo boli konkrétne texty označené ako netypické/anomálne (Obrázok 1), pričom sa analyzuje vplyv jednotlivých atribútov a ich hodnôt na výsledok modelu pre daný text. Zároveň je možné získať aj globálny pohľad na správanie modelu, napríklad prostredníctvom beeswarm (Obrázok 3) vizualizácií, ktoré zobrazujú význam a distribúciu vplyvu jednotlivých atribútov naprieč celou dátovou sadou.

Navrhnutá metóda bola experimentálne overená na piatich dátových sadách súdnych rozhodnutí, ktoré boli manuálne anotované právnikmi. Každá dátová sada obsahuje súdne rozhodnutia z vybranej oblasti práva, ako napríklad „zmluvná pokuta“ či „nekalá súťaž“, čím bolo možné overiť robustnosť navrhnutého prístupu naprieč rôznymi typmi oblastí. Metóda pozostáva z troch hlavných krokov: prípravy a transformácie vstupných dát do numerickej reprezentácie, detekcie netypických textov pomocou modelu Isolation Forest a následnej interpretácie výsledkov pomocou metódy SHAP.

Týmto spôsobom metóda pokrýva celý proces od spracovania právnych atribútov, cez identifikáciu netypických textov, až po vysvetlenie dôvodov, prečo boli konkrétne texty modelom označené ako netypické.

1. Príprava dát a úprava atribútov

Pred aplikáciou modelu boli vstupné dáta transformované do numerickej reprezentácie vhodnej pre spracovanie modelom Isolation Forest. Právne dokumenty boli reprezentované prostredníctvom štruktúrovaných atribútov získaných z anotovaných dátových sád.

Proces predspracovania zahŕňal:

- spracovanie numerických atribútov vrátane doplnenia chýbajúcich hodnôt a vytvorenia indikátorov chýbajúcich hodnôt,
- normalizáciu textových atribútov (prevod na malé písmená, odstránenie diakritiky),
- transformáciu binárnych kategórií (napr. „áno“ / „nie“) do numerickej podoby,
- doplnenie hodnoty „unknown“ pre chýbajúce kategórie,
- transformáciu kategorických atribútov pomocou one-hot encodingu,
- odstránenie atribútov s konštantnou hodnotou.

Výsledkom tohto procesu bola numerická dátová sada, ktorá bola následne použitá ako vstup pre model detekcie outlierov.

2. Model pre detekciu outlierov

Na detekciu netypických právnych rozhodnutí bol použitý model Isolation Forest, ktorý patrí medzi metódy učenia bez učiteľa.

Tento model funguje na princípe náhodného rozdeľovania dátového priestoru. Rozhodnutia, ktoré sa dajú izolovať menším počtom rozdelení na základe ich atribútov, sú považované za netypické. Dôvodom je, že takéto rozhodnutia majú hodnoty atribútov, ktoré sa výrazne líšia od väčšiny ostatných – napríklad obsahujú neobvyklé kombinácie atribútov alebo extrémne hodnoty. Pri náhodnom delení dát preto dochádza k ich rýchlejšiemu oddeleniu od zvyšku dátovej sady, čo vedie k ich označeniu ako netypických. Naopak, rozhodnutia, ktoré zdieľajú podobné charakteristiky s väčšinou dát, vyžadujú viacero rozdelení na ich izolovanie, a preto sú považované za bežné.

Takýto prístup je vhodný pre situácie, kde nie sú k dispozícii explicitné anotácie outlierov.

Model bol aplikovaný na jednotlivé dátové sady samostatne a pre každé rozhodnutie boli vypočítané:

- **skóre anomálnosti**, ktoré vyjadruje mieru odlišnosti rozhodnutia od ostatných na základe jeho atribútov. Toto skóre je odvodené z toho, ako rýchlo je dané rozhodnutie izolované v rámci modelu – rozhodnutia, ktoré sú izolované rýchlejšie, dosahujú nižšie hodnoty skóre a sú považované za viac netypické,
- **indikátor označujúci, či ide o netypické rozhodnutie**, ktorý vzniká na základe porovnania skóre s interným prahom modelu (daným parametrom kontaminácie, v našom prípade 10%),
- **poradie rozhodnutí podľa miery odchýlky**, ktoré umožňuje zoradiť rozhodnutia od najviac netypických po najviac typické.

3. Interpretácia výsledkov pomocou SHAP

Na interpretáciu výsledkov modelu bola použitá metóda SHAP (SHapley Additive exPlanations), ktorá umožňuje vysvetliť rozhodnutia modelu na úrovni jednotlivých atribútov.

Pomocou SHAP je možné:

- vysvetliť, prečo bolo **konkrétne súdne rozhodnutie** označené ako anomálne,
- určiť, ktoré atribúty a ich konkrétne hodnoty najviac prispeli k tomuto rozhodnutiu,
- analyzovať smer vplyvu jednotlivých atribútov (či prispievajú k netypickosti alebo naopak k „typickosti" prípadu).

Pre každý dokument je možné zobraziť detailné vysvetlenie vo forme tzv. waterfall grafov, ktoré znázorňujú príspevok jednotlivých atribútov k výslednému skóre.

Na obrázku Obrázok 1 je zobrazený tzv. waterfall graf pre konkrétne rozhodnutie. Tento graf znázorňuje, ako jednotlivé atribúty a ich hodnoty postupne ovplyvňujú výsledné skóre modelu.

Základom grafu je priemerná hodnota modelu (baseline), od ktorej sa postupne pripočítavajú alebo odpočítavajú príspevky jednotlivých atribútov. Každý riadok predstavuje jeden atribút spolu s jeho konkrétnou hodnotou pre dané súdne rozhodnutie.

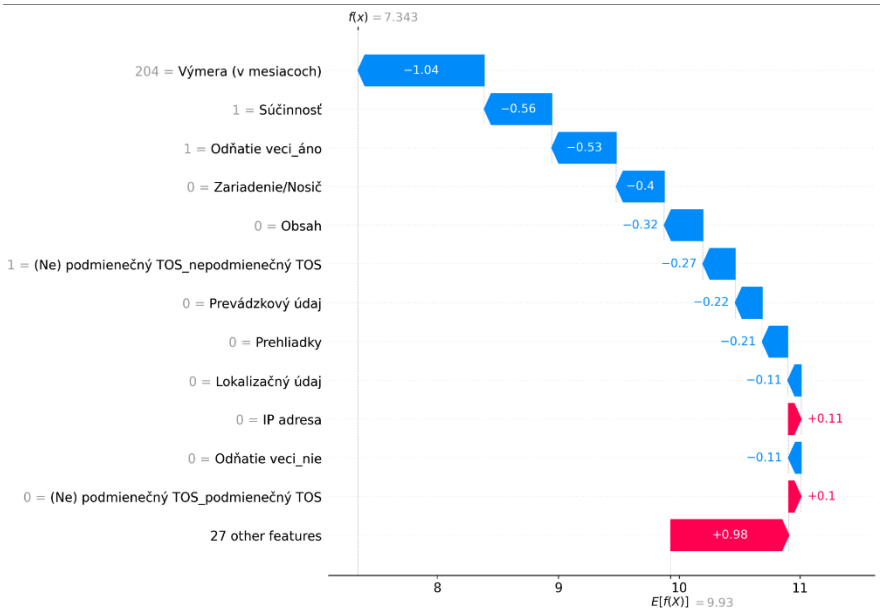
Pri modeli Isolation Forest je dôležité, že nižšie výsledné skóre znamená vyššiu mieru anomálnosti. Preto atribúty s negatívnym príspevkom znižujú výsledné skóre modelu pre dané súdne rozhodnutie, čím ho posúvajú smerom k netypickým prípadom, zatiaľ čo atribúty s pozitívnym príspevkom toto skóre zvyšujú a prispievajú k jeho zaradeniu medzi bežnejšie rozhodnutia.

Výsledná hodnota grafu f(x) predstavuje finálne skóre modelu pre dané rozhodnutie. Z grafu je tak možné identifikovať, ktoré atribúty najviac prispeli k tomu, že rozhodnutie bolo vyhodnotené ako netypické, a ktoré atribúty tento efekt naopak zmierňovali.

Z grafu je možné detailne interpretovať aj konkrétne atribúty. Napríklad:

- Atribút **„Výmera (v mesiacoch) = 204"** má výrazný negatívny príspevok (−1.04), čo znamená, že jeho hodnota výrazne znižuje výsledné skóre modelu pre dané súdne rozhodnutie a tým ho posúva smerom k netypickým prípadom. Na rozdiel od binárnych atribútov ide v tomto prípade o numerický atribút, pričom konkrétna hodnota (204 mesiacov) predstavuje extrémnejšiu hodnotu v rámci dátovej sady, čo vedie k jeho výraznému vplyvu na výsledok modelu. Ide zároveň o atribút s najväčším vplyvom na výsledok modelu pre toto rozhodnutie.

Tieto hodnoty ukazujú, že kombinácia konkrétnych atribútov a ich hodnôt môže výrazne ovplyvniť výsledok modelu a viesť k označeniu rozhodnutia ako anomálneho.



Obrázok 1 - Vysvetlenie anomálnosti konkrétneho súdneho rozhodnutia pomocou waterfall grafu

Na globálnej úrovni boli využité:

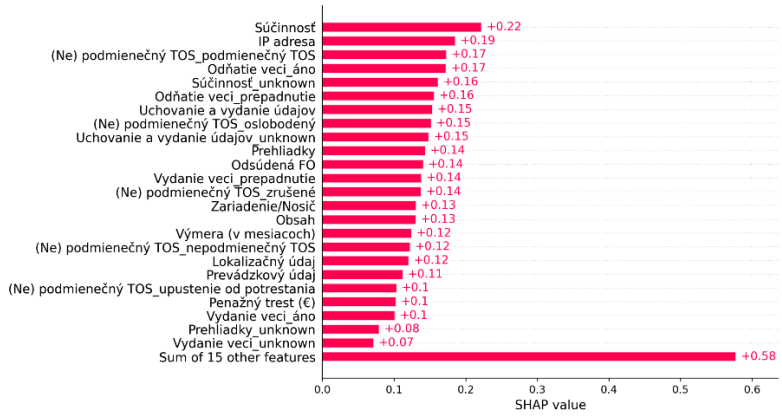
- graf dôležitosti atribútov (Obrázok 2),
- beeswarm vizualizácia, ktorá zobrazuje distribúciu vplyvu atribútov naprieč celou dátovou sadou (Obrázok 3),
- analýza interakcií medzi atribútmi pomocou SHAP interaction hodnôt (Obrázok 4).

Na ďalšom obrázku Obrázok 2 je znázornená globálna dôležitosť jednotlivých atribútov pomocou priemernej absolútnej hodnoty SHAP (mean absolute SHAP values).

Tento typ vizualizácie poskytuje prehľad o tom, ktoré atribúty majú najväčší vplyv na rozhodovanie modelu naprieč celou dátovou sadou.

Každý riadok predstavuje jeden atribút a hodnota na osi X (SHAP value) vyjadruje priemerný absolútny vplyv daného atribútu na výsledné skóre modelu. Čím je hodnota vyššia, tým väčší vplyv má daný atribút na rozhodovanie modelu, bez ohľadu na to, či tento vplyv zvyšuje alebo znižuje skóre anomálnosti.

Je dôležité zdôrazniť, že tento graf neukazuje smer vplyvu (t. j. či atribút zvyšuje alebo znižuje anomálnosť), ale iba jeho veľkosť.



Obrázok 2 - Globálna dôležitosť atribútov vyjadrená pomocou priemernej absolútnej hodnoty SHAP, zobrazujúca atribúty s najväčším vplyvom na rozhodovanie modelu.

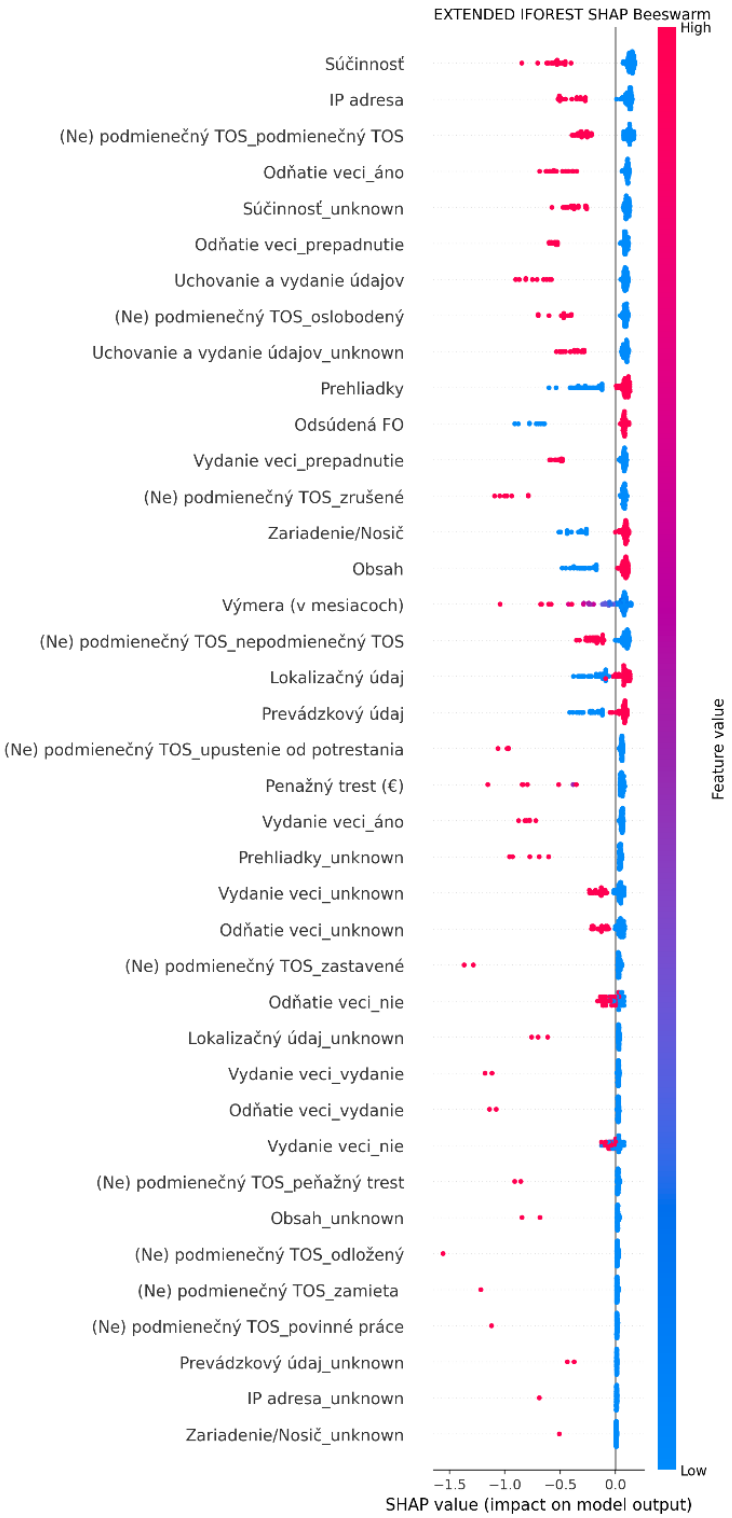
Na ďalšom obrázku Obrázok 3 je znázornená tzv. beeswarm vizualizácia SHAP hodnôt, ktorá poskytuje detailnejší pohľad na vplyv jednotlivých atribútov naprieč celou dátovou sadou.

Každý riadok predstavuje jeden atribút a každý bod zodpovedá jednému konkrétnemu súdnemu rozhodnutiu. V každom riadku sa nachádza toľko bodov, koľko je rozhodnutí v danej dátovej sade. Poloha bodu na osi X vyjadruje SHAP hodnotu, teda vplyv daného atribútu na výsledné skóre modelu pre konkrétne rozhodnutie. V prípade modelu Isolation Forest platí, že hodnoty posunuté viac doľava znižujú výsledné skóre (teda prispievajú k anomálnosti), zatiaľ čo hodnoty bližšie k nule alebo viac vpravo prispievajú k bežnejším rozhodnutiam.

Farba bodu reprezentuje hodnotu atribútu – modrá farba zodpovedá nižším hodnotám atribútu, zatiaľ čo červená farba reprezentuje vyššie hodnoty. Kombináciou polohy a farby bodov je možné identifikovať, ako konkrétne hodnoty atribútov ovplyvňujú výsledok modelu.

Z grafu je zároveň možné pozorovať rozdiely medzi typmi atribútov. Niektoré atribúty majú iba dve hodnoty (napr. binárne atribúty typu áno/nie), čo sa prejavuje tým, že body sú zobrazené len v dvoch farbách – modrej a červenej. Naopak, atribúty s plynulými alebo numerickými hodnotami (napr. „Výmera (v mesiacoch)“) vytvárajú farebné spektrum od modrej po červenú, čo odráža širší rozsah možných hodnôt a ich postupný vplyv na výsledné skóre modelu.

Z grafu je možné interpretovať aj konkrétne atribúty. Napríklad pri atribúte **„Súčinnosť“** je vidieť, že rozhodnutia s hodnotou 1 (áno) sú systematicky posunuté doľava, teda prispievajú k anomálnosti. Naopak, pri atribúte **„Prehliadky“** platí opačný efekt – rozhodnutia s hodnotou 0 majú tendenciu prispievať k anomálnosti, čo naznačuje, že prítomnosť tohto atribútu je v dátovej sade skôr typická.



Obrázok 3 - Beeswarm vizualizácia SHAP hodnôt zobrazujúca vplyv atribútov na výsledné skóre modelu naprieč všetkými rozhodnutiami.

Na ďalšom obrázku Obrázok 4 je znázornená heatmapa interakcií medzi atribútmi na základe SHAP hodnôt. Táto vizualizácia zachytáva, do akej miery dvojice atribútov spoločne ovplyvňujú výsledné skóre modelu.

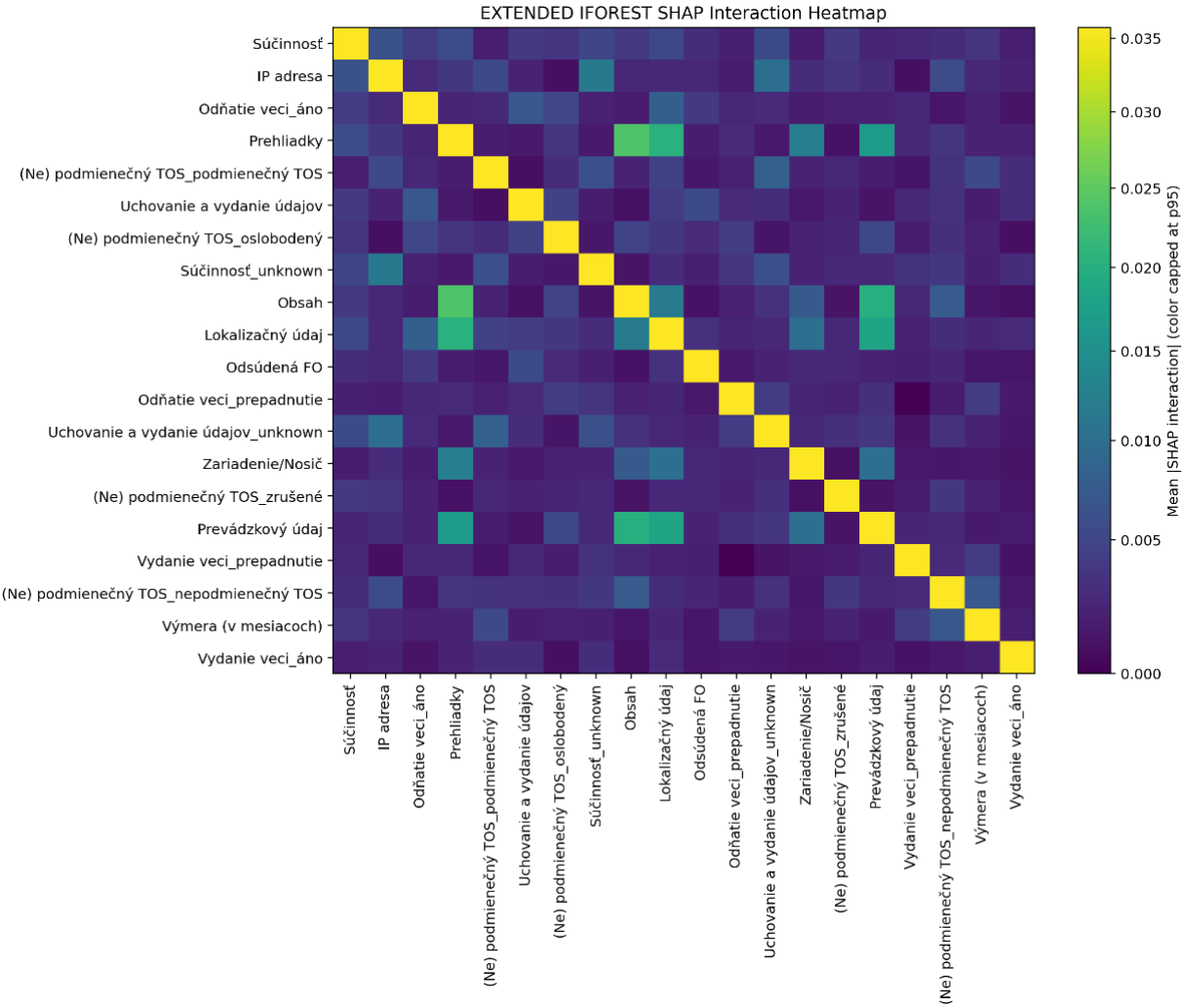
Každý riadok a stĺpec reprezentuje jeden atribút a hodnota v bunke vyjadruje priemernú absolútnu silu interakcie medzi danou dvojicou atribútov naprieč celou dátovou

sadou. Použitie absolútnej hodnoty zabezpečuje, že sa zohľadňuje celková intenzita interakcie bez ohľadu na jej smer (t. j. či interakcia zvyšuje alebo znižuje výsledné skóre modelu). Bez použitia absolútnej hodnoty by sa pozitívne a negatívne interakcie mohli navzájom rušiť a skryť tak skutočne významné vzťahy medzi atribútmi. Čím je hodnota vyššia (svetlejšia farba), tým silnejší je spoločný vplyv týchto atribútov na rozhodovanie modelu.

Hodnoty na diagonále predstavujú individuálny (hlavný) vplyv jednotlivých atribútov, zatiaľ čo hodnoty mimo diagonály zachytávajú skutočné interakcie medzi rôznymi atribútmi.

Z grafu je možné pozorovať, že väčšina interakcií medzi atribútmi je relatívne slabá, pričom dominantnú úlohu zohrávajú skôr individuálne atribúty. Silnejšie interakcie (svetlejšie oblasti mimo diagonály) indikujú dvojice atribútov, ktorých kombinácia má významnejší vplyv na výsledok modelu než každý z nich samostatne.

Táto analýza umožňuje identifikovať nielen dôležité atribúty, ale aj ich vzájomné vzťahy, ktoré môžu byť kľúčové pre pochopenie správania modelu.



Obrázok 4 - Heatmapa SHAP interakcií zobrazujúca silu vzájomného vplyvu medzi dvojicami atribútov na výsledné skóre modelu.

Tieto prístupy umožňujú lepšie pochopiť správanie modelu a identifikovať hlavné faktory vedúce k označeniu dokumentov ako anomálnych, pričom zároveň poskytujú transparentné a interpretovateľné vysvetlenie rozhodnutí modelu na úrovni jednotlivých rozhodnutí aj celého datasetu.

4. Navrhnutá metóda a jej výstupy

Výsledkom riešenia je metóda na identifikáciu a interpretáciu netypických právnych rozhodnutí, ktorá bola aplikovaná na päť dátových sád.

Pre každú dátovú sadu sa vytvára samostatný výstupný priečinok (napr. *ALTAML_{dataset}_anomaly_detection/iforest*), kde *{dataset}* predstavuje názov konkrétnej dátovej sady), obsahujúci kompletné výsledky detekcie a ich interpretácie.

Navrhnutý prístup nepredstavuje iba priamu aplikáciu modelu Isolation Forest, ale zahŕňa aj spracovanie dát, transformáciu právnych atribútov do numerickej reprezentácie a integráciu metód vysvetliteľnosti (SHAP), ktoré umožňujú interpretovať výsledky v kontexte právnych rozhodnutí.

Konkrétne výstupy sú:

- **{dataset}_rows_with_iforest_scores.csv**
  obsahuje pôvodné rozhodnutia doplnené o skóre anomálnosti, označenie netypických rozhodnutí (outlierov) a poradie podľa miery odchýlky.
- **{dataset}_iforest_shap_values.csv**
  obsahuje SHAP hodnoty pre jednotlivé atribúty a rozhodnutia, teda lokálne číselné vysvetlenia modelu.
- **{dataset}_iforest_global_shap_importance.csv**
  obsahuje globálnu dôležitosť atribútov vypočítanú ako priemerná absolútna SHAP hodnota.
- **{dataset}_iforest_shap_bar_mean_abs.png**
  vizualizuje najdôležitejšie atribúty podľa priemernej absolútnej SHAP hodnoty (Obrázok 2).
- **{dataset}_iforest_shap_beeswarm.png**
  zobrazuje globálne správanie modelu a smer vplyvu atribútov na rozhodnutia (Obrázok 3).
- **waterfalls/**
  obsahuje SHAP waterfall graf pre každé rozhodnutie, teda detailné lokálne vysvetlenie konkrétneho prípadu (Obrázok 1).
- **{dataset}_iforest_interaction_pairs.csv**
  obsahuje zoradené dvojice atribútov podľa sily ich SHAP interakcie.
- **{dataset}_iforest_interaction_heatmap.png**
  vizualizuje najsilnejšie interakcie medzi atribútmi (Obrázok 4).
- **{dataset}_iforest_top_pair_per_row.csv**
  obsahuje najvýznamnejšiu interakčnú dvojicu pre najviac anomálne rozhodnutia.
- **detector_summary.csv**
  sumarizuje výsledky pre konkrétny dataset (napr. počet rozhodnutí, počet atribútov, počet označených netypických rozhodnutí).

Okrem toho sa vytvára aj spoločný súbor **all_datasets_detector_summary.csv**, ktorý sumarizuje výsledky naprieč všetkými spracovanými datasetmi.

Navrhnutá metóda je zároveň flexibilná a umožňuje analyzovať rôzne aspekty právnych rozhodnutí vrátane ich výsledkov (verdiktov), ktoré sú reprezentované ako atribúty v dátach. Vďaka tomu je možné identifikovať netypické rozhodnutia nielen na základe jednotlivých atribútov, ale aj na základe ich kombinácií.

5. Zdrojové súbory

Zdrojový kód projektu je spravovaný v internom repozitári UPJŠ GitLab. Je dostupný na nasledujúcej adrese https://gitlab.science.upjs.sk/altaml/method-for-outlier-detection a obsahuje aj spustiteľnú implementáciu navrhnutej metódy nad piatimi použitými dátovými sadami.