

Téma 1: PREDIKCE VÝSKYTU HOREČKY DENGUE

Motivace

Pro různá místa na světě jsou typické i výskyty specifických nemocí. Horečka dengue je jednou z nemocí přenášenou převážně hmyzem v oblastech Karibiku, Filipín, Malajsie, Tchaj-wanu nebo Indie. Do ostatních regionů se tato nemoc dostává kvůli neustále se zvyšujícímu počtu turistů. V dnešní době neexistuje účinná vakcína, léčba je pouze symptomatická. Ačkoliv většina nemocných se vyléčí bez jakýchkoliv následků, u některých skupin pacientů (děti, osoby s chronickými nemocemi) může dengue vést ke vzniku závažných zdravotních komplikací (vnitřní krvácení, šokový syndrom). Predikce výskytu dengue napomáhá např. včasnému zajištění bezpečnostních opatření při cestování do daného regionu nebo doporučení pro místní obyvatelstvo. Vzhledem ke způsobu přenosu nemoci (komáři), predikci jejího výskytu lze úspěšně provést za použití meteorologických, sociodemografických aj. dat.

Cíl projektu

Cílem je predikovat **incidenci horečky dengue** s co nejvyšší přesností na základě dostupných meteorologických dat.

Datové soubory

Data pochází z oblasti San-Juan (Portoriko). K dispozici jsou různé meteorologické parametry měřené v dané lokalitě a údaje o výskytu horečky dengue v místní populaci. Meteorologické parametry mohou sloužit jako příznaky pro trénování predikčního modelu. Údaje o výskytu dengue v místní populaci mohou sloužit jako očekávané výstupní hodnoty modelu. Jelikož se jedná o predikční úlohu, pro trénování modelu by se měla použít pouze data z období předcházejícího okamžiku predikce, tj. např. pro predikci dengue v 5. týdnu sledování lze využít pouze data z předchozího období (tj. z 1.-4. týdne). Délka sledovaného a analyzovaného časového okna přitom musí respektovat známé souvislosti mezi predikovanou hodnotou a použitými příznaky. Pokud se jedná o včasnou predikci dengue s předstihem 2 týdny dopředu, pak pro predikci např. v 5. týdnu sezóny lze pro trénování využít pouze data z období 1.-2. týden.

SanJuanData.csv – shrnuje meteorologické údaje a informaci o výskytu horečky dengue v dané lokalitě.

Unikátní meteorologická data pocházející z globálního celosvětového měření různých veličin. Výsledné parametry jsou získány v rámci komplexního zpracování a analýzy všech nasbíraných údajů. Data jsou sbírána s frekvencí 1x denně po dobu několika kalendářních let.

Údaj o výskytu horečky dengue je vyjádřen jako celkové množství pacientů s diagnostikovanou horečkou dengue. Tento údaj zahrnuje pacienty s pozitivním PCR nálezem pro jeden ze čtyř antigeně odlišných sérotypů dengue (DENV1, DENV2, DENV3, DENV4 – „přímá“ diagnostika přítomnosti samotného viru dengue) a pacienty s pozitivním sérologickým vyšetřením („nepřímá“ diagnostika dengue - detekce přítomnosti protilátek třídy IgM a IgG produkovaných organismem v případě kontaktu pacienta s virem). Sběr těchto dat se provádí v tzv. sezóně sledování dengue, která začíná týdnem s nejnižší incidencí nemoci v dané lokalitě (zjištění za celou dobu sledování dengue v dané lokalitě) a končí po uplynutí 12 měsíců. Začátek sezóny sledování nemoci nekoresponduje se začátkem kalendářního roku. Hodnoty jsou zaznamenávány 1x týdně.

Year: rok měření
Month: měsíc měření
Day: pořadí dne
airTemp: teplota vzduchu (K)
dewPoint: teplota rosného bodu (K)
humidityRelative: relativní vlhkost vzduchu (%)
humiditySpecific: měrná vlhkost vzduchu (g kg⁻¹)
tempMax: maximální denní teplota vzduchu (K)
tempMin: minimální denní teplota vzduchu (K)
tempAvg: průměrná denní teplota vzduchu (K)
tempMinMaxDiff: denní rozpětí mezi minimální a maximální teplotou (K)
Precipitation: denní množství atmosférických srážek (kg m⁻²)
dengueCases: celkové množství pacientů s diagnostikovanou horečkou dengue (**použijte pro odvození očekávaných hodnot incidence onemocnění**)
populationTotal: odhadovaná velikost sledované populace v daném roce

POZOR:

Chybějící hodnoty měřených veličin jsou označeny: NaN

Výstup projektu

Výstupem projektu budou hodnoty **incidence dengue pro každý sledovaný týden s předstihem 1 měsíce**. Pro vyhodnocení úspěšnosti predikce použijte *průměrnou absolutní odchylku* (mean absolute error, MAE) mezi predikovanou hodnotou incidence y a očekávanou hodnotou incidence d (n – počet predikovaných hodnot):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - d_i|$$

Další informace o horečce dengue např. zde: http://www.khsstc.cz/dokumenty/horecka-dengue-4658_4658_161_1.html

Téma 2: DETEKCE SEPSE NA ZÁKLADĚ KLINICKÝCH DAT

Motivace

Sepse je život ohrožující stav, který je v českém prostředí často označován nepřesně jako “otrava krve”. Reálně je ale sepsa jednou z nejzávažnějších a nejčastějších komplikací v traumatologii, chirurgii a obecně intenzivní péči. Jedná se o celkovou reakci organismu na infekci, může přecházet v septický šok zahrnující i orgánové selhání. Mortalita spojená se sepsí v posledních letech “porazila” i mortalitu na infarkt myokardu. Náklady na léčbu těžké sepsy se pohybují v řádech statisíců Kč. Vzhledem k oběma uvedeným okolnostem je nasnadě automaticky detekovat nástup septického stavu a tím snížit nároky na personál JIP a zároveň eliminovat lidský faktor. Septický stav je možné detekovat na základě sledování vitálních funkcí pacienta (krevní tlak, teplota, tep, ...) nebo laboratorních vyšetření. Pomoci mohou i demografické charakteristiky (věk, pohlaví,...).

Cíl projektu

Cílem je určit, **zda se pacient nachází v septickém stavu či nikoli** na základě dostupných klinických dat.

Datové soubory

dataSepsis.csv – klinická data pacientů JIP a údaje o stavu pacienta.

Vitální charakteristiky, výsledky laboratorního vyšetření a demografické charakteristiky mohou sloužit k trénování klasifikačního modelu za účelem hodnocení stavu pacienta (septický či neseptický). Jednotlivé hodnoty se sbírají v různých časech s rozdílnými intervaly mezi měřeními. Z tohoto důvodu může v záznamech část údajů chybět.

Vitální charakteristiky

HR	Tepová frekvence (bpm)
O2Sat	Saturace O2 (%)
Temp	Teplota těla (° C)
SBP	Systolický tlak (mmHg)
MAP	Střední arteriální tlak (mmHg)
DBP	Diastolický tlak (mmHg)
Resp	Dechová frekvence (počet dechů za minutu)
EtCO2	Obsah CO2 ve vzduchu na konci výdechu (end-tidal CO2) (mmHg)

Laboratorní vyšetření

BaseExcess	Measure of excess bicarbonate (mmol/L)
HCO3	Bikarbonáty (mmol/L)
FiO2	koncentrace O2 ve vdechovaném vzduchu (%)
pH	N/A

PaCO2	Parciální tlak CO2 v arteriální krvi (mmHg)
SaO2	Saturace O2 v arteriální krvi (%)
AST	Aspartátaminotransferáza (IU/L)
BUN	Močovinový dusík v krvi (mg/dL)
Alkalinephos	Alkalická fosfatáza (IU/L)
Calcium	Vápník (mg/dL)
Chloride	Chloridy (mmol/L)
Creatinine	Kreatinin (mg/dL)
Bilirubin_direct	Bilirubin přímý (mg/dL)
Glucose	Glykemie (mg/dL)
Lactate	Laktát (mg/dL)
Magnesium	Hořčík (mmol/dL)
Phosphate	Fosfát (mg/dL)
Potassium	Draslík (mmol/L)
Bilirubin_total	Bilirubin celkový (mg/dL)
TroponinI	Troponin I (ng/mL)
Hct	Hematokrit (%)
Hgb	Hemoglobin (g/dL)
PTT	Aktivovaný parciální tromboplastinový čas (s)
WBC	Leukocyty (*10 ³ /μL)
Fibrinogen	Fibrinogen (mg/dL)
Platelets	Trombocyty (count*10 ³ /μL)

Demografické charakteristiky

Age	Věk
Gender	Pohlaví: žena (0) or muž (1)
Unit1	ID JIP (MICU)
Unit2	ID JIP (SICU)
HospAdmTime	Počet hodin mezi hospitalizací a umístěním na JIP
ICULOS	Doba strávená na JIP (hod)

Stav pacienta

isSepsis	0 – neseptický, 1 – septický (odhadovaná veličina)
----------	---

POZOR:

V datovém souboru se mohou vyskytovat odlehlé hodnoty, chybějící hodnoty měřených veličin (označené jako např. NaN, '-' apod.), hodnoty nedávající smysl apod.

Výstup projektu

Výstupem projektu budou hodnoty (0/1) indikující stav pacienta. Pro vyhodnocení úspěšnosti klasifikace dat použijte matici záměn (confusion matrix) a Se, Sp, Acc (v případě rovnoměrně zastoupených klasifikačních skupin) či F-measure (v případě nerovnoměrně zastoupených klasifikačních skupin):

$$Se = TP / (TP + FN)$$

$$Sp = TN / (TN + FP)$$

$$Acc = (TP + TN) / (TP + TN + FP + FN)$$

$$PPV = TP / (TP + FP)$$

$$F = 2 \frac{PPV * Se}{PPV + Se}$$

Téma 3: ROZPOZNÁNÍ ČÍSLIC Z OBRAZOVÝCH DAT (OPTICAL CHARACTER RECOGNITION, OCR)

Motivace

Rozpoznání textu je jednou ze základních úloh automatického zpracování obrazů s výskytem textových dat. Typické využití lze nalézt např. při rozpoznávání poznávacích značek automobilů ze silničních kamer, strojovém čtení archivních lékařských zpráv, strojové rozpoznávání čísel hráčů na dresech či u pomůcek pro nevidomé. Algoritmus by měl být schopen rozpoznat v obrazu znak za různě ztížených podmínek (snížená viditelnost, geometrické zkreslení, šum, obrazové artefakty) a bez ohledu na typ písma, kterým byl znak napsán.

Cíl projektu

Cílem projektu je vytvořit klasifikační model, který ve vstupním obrazu **rozpozná číslice z intervalu 0 – 9** a přiřadí jim odpovídající hodnotu dle zadání.

Datové soubory











Data pro trénování a testování modelu jsou dostupná ve složce trainData. Finální testování bude provedeno na skryté sadě dat. Udržujte proto strukturu dat, která vám byla poskytnuta (v opačném případě nebude možné projekt automaticky vyhodnotit a bude hodnocen 0 body).

characterName_id.png – Obrazová data znaku ve formátu *.png (barevná hloubka 8 bitů, 3 kanály).

references.csv – Tabulka obsahující referenční hodnoty (požadované třídy) pro jednotlivé obrázky.

Název třídy odpovídá danému znaku (např. znak „3“ je zařazen do třídy 3). Výjimku tvoří znak „0“, který je klasifikován do třídy označené 10.

Ukázka dat:

Vstupní data:										
	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
Výstup:	10	3	7	5	2	8	6	9	4	1

Výstup projektu

Pro vyhodnocení úspěšnosti klasifikace dat použijte matici záměn (confusion matrix) a F1 skóre.

$$F = 2 \frac{PPV * Se}{PPV + Se}$$

Téma 4: PREDIKCE DÉLKY KARIÉRY HRÁČŮ V NBA

Motivace

Špičkové sportovní ligy po celém světě představují na jednu stranu prostředek zábavy pro široké masy lidí, na druhou stranu je to ale business model, ve kterém se točí miliardy. Pro výdělek jednotlivé kluby podnikají investice v podobě nákupu hráčů a výdělky inkasují ve formě reklamy nebo prodeje vysílacích práv televizním společnostem. Pro efektivitu takového modelu je nutné umět se dobře rozhodnout, která investice je dostatečně vratná, tudíž vědět, jaký potenciál mají hráči a jak prestižní daný tým díky nim bude. Umět predikovat z dostupných dat sbíraných o hráčích jejich působení i na další roky by znamenalo pro týmy jistější investice a pro diváky zajímavější akce a intenzivnější zážitky. Jednou ze špičkových světových lig je i americká NBA. Z dostupných statistických dat je někdy možné předpovědět, zda hráčova kariéra v NBA potrvá více nebo méně než 5 let. Pro výběr vhodného hráče jsou potom důležité různé parametry popisující jeho výkon v první sezoně NBA.

Cíl projektu

Cílem je určit (odhadnout), **zda sportovec bude působit v NBA déle než 5 let či nikoliv**.

Datové soubory

dataCareer.csv – statistické údaje o sportovcích vypočítané na základě odehraných zápasů.

Individuální statistické údaje ze zápasů odehraných v poslední době (viz dole) mohou sloužit k trénování klasifikačního modelu za účelem odhadu úspěšnosti sportovní kariéry sportovce v rámci NBA (z hlediska její délky).

Označení příznaku v databázi	Popis příznaku
Var1	Pořadí sportovce v databázi
Name	Jméno sportovce
GP	Počet odehraných zápasů
MIN	Počet odehraných minut
PTS	Počet bodů za hru
FGM	Proměněné střelecké pokusy
FGA	Střelecké pokusy
FG	Střelecké pokusy (%)
x3P_Made	Třibodové hody - proměněné
x3PA	Třibodové hody - pokusy
x3P	Třibodové hody (%)
FTM	Trestné hody - proměněné
FTA	Trestné hody - pokusy
FT	Trestné hody (%)
OREB	Útočné doskoky

DREB	Obranné doskoky
REB	Doskoky
AST	Přihrávky
STL	Zisk míče
BLK	Bloky
TOV	Ztráty míče v útoku

Výstup predikce (tzv. požadovaná hodnota) je obsažena v posledním (22.) sloupci datové matice pod proměnnou s názvem `Target_5YRS`. Tento příznak může nabývat hodnoty **‘yes’** – působení v NBA pro daného hráče je delší než 5 let nebo **‘no’** – kariéra v NBA kratší než 5 let.

POZOR:

- V datovém souboru se mohou vyskytovat odlehlé hodnoty, chybějící hodnoty měřených veličin (označené jako např. NaN, ‘-’ apod.), hodnoty nedávající smysl apod.
- Pro hladké fungování kódu je nutné vhodným způsobem transformovat (číselně zakódovat) požadované výstupní hodnoty **‘yes/no’** (viz info níže a v kódech přiložených k projektu).

Výstup projektu

Výstupem projektu budou hodnoty (0/1) odpovídající odhadované délce působení konkrétních sportovců v NBA, kde 0 odpovídá krátkodobému působení (< 5 let) a 1 odpovídá dlouhodobějšímu působení v NBA (> 5 let, zřejmě spojené s úspěšným nasazením sportovce). Pro vyhodnocení úspěšnosti klasifikace dat použijte *matici záměn* (confusion matrix) a *Se*, *Sp*, *Acc* (v případě rovnoměrně zastoupených klasifikačních skupin) či *F-measure* (v případě nerovnoměrně zastoupených klasifikačních skupin), které jsou implementované v přiloženém skriptu *GetScoreCareer*:

$$Se = TP / (TP + FN)$$

$$Sp = TN / (TN + FP)$$

$$Acc = (TP + TN) / (TP + TN + FP + FN)$$

$$PPV = TP / (TP + FP)$$

$$F = 2 \frac{PPV * Se}{PPV + Se}$$

Téma 5: AUTOMATICKÁ PREDIKCE STŘELECKÉ ÚSPĚŠNOSTI

Motivace

V poslední době začíná růst význam moderních technologií, včetně umělé inteligence, ve sportu. Fotbal byl dlouho sportem, který byl v tomto směru na druhé koleji. Nicméně i to se v posledních letech začíná měnit. Důležitým parametrem v zápasech je střelecká úspěšnost. Ta se odvíjí od mnoha věcí, jako je například pozice odkud hráč střílí (malé vápno nebo středový kruh), situace, ze které střela vzešla (sólo nebo rohový kop) nebo část těla, kterou byla v kontaktu s míčem jako poslední (noha nebo hlavička). Data, se kterými budete pracovat, jsou sbíraná v průběhu lig 2014/2015 až 2019/2020, a to u pěti prestižních fotbalových lig (English Premier League, La Liga, Bundesliga, Serie A, Ligue 1). Výsledky predikcí by mohly vést k lepšímu plánování sestav v zápasech, změně rozložení sil během zápasu, zpětné analýzy chyb nebo k upřesnění kurzů na jednotlivé hráče nebo výsledky zápasů v sázkových kancelářích.

Cíl projektu

Cílem je vytvoření klasifikačního modelu, který zvládne vstupním datům přiřadit jednu z kategorií 'BlockedShot', 'Goal', 'MissedShots', 'OwnGoal', 'SavedShot' nebo 'ShotOnPost'.

Datové soubory

dataShots.csv – údaje popisující samotnou ránu na branku, situaci, která tomu předcházela, apod.

Jedná se celkem o různých 19 příznaků měřených přímo v souvislosti s daným pokusem o gól.

Označení příznaku v databázi	Popis příznaku	Poznámka
<code>id</code>	ID zápasu	
<code>minute</code>	minuta, ve které padla střela	
<code>X</code>	pokud velikost fotbalového hřiště je 100x100 jednotek (délka x šířka, zjednodušená reprezentace), pak X reprezentuje délku (resp. první souřadnici místa, odkud vycházela střela). 0 až 100 jednotek zleva doprava mezi brankovými čarami	
<code>Y</code>	Y reprezentuje šířku (resp. druhou souřadnici místa, odkud vycházela střela). 0 až 100 mezi postranními čarami (od outu k outu)	
<code>xG</code>	xG (=Expected Goal) dané střely	
<code>player</code>	hráč-střelec	
<code>h_a</code>	jestli střela padla doma (v domácím zápase) nebo ne	
<code>player_id</code>	ID hráče-střelce	

situation	situace, ze které se střílelo	např. OpenPlay, SetPiece, DirectFreekick, FromCorner
year	daná sezóna	
shotType	část těla, kterou hráč vystřelil	např. RightFoot, Head, LeftFoot, Other
match_id	ID daného zápasu	
h_team	domácí tým	
a_team	hostující tým	
h_goals	góly/body domácího týmu	
a_goals	góly/body hostujícího týmu	
date	datum zápasu	
player_assisted	přihrávající hráč (pokud byl)	
lastAction	akce předcházející střele	např. Chipped, Cross, Pass, TakeOn, Rebound

Výstup predikce (tj. požadovaný výstup) lze najít v datové matici ve 3. sloupci pod proměnnou s názvem `Result`. Tato proměnná může nabývat hodnot: **'BlockedShot', 'Goal', 'MissedShots', 'OwnGoal', 'SavedShot'** nebo **'ShotOnPost'**. Jedná se tedy o tzv. vícekategoriální klasifikaci (angl. multiclass classification), kdy se pracuje s více než dvěma klasifikačními třídami.

POZOR:

- V datovém souboru se mohou vyskytovat odlehlé hodnoty, chybějící hodnoty měřených veličin (označené jako např. NaN, '-' apod.), hodnoty nedávající smysl apod.
- Pro hladké fungování kódu je nutné vhodným způsobem transformovat (číselně zakódovat) požadované výstupní hodnoty (viz info v kódech přiložených k projektu).

Výstup projektu

Pro vyhodnocení úspěšnosti klasifikace dat použijte matici záměn (confusion matrix) a F1 skóre, které jsou implementované v přiloženém skriptu `GetScoreShots`.

$$F = 2 \frac{PPV * Se}{PPV + Se}$$

Téma 6: PREDIKCE POČTU MEDAILÍ NA OH

Motivace

Zatímco někteří sportovci jezdí na olympiády roky a medaile vidí jen u svých kolegů, jsou i tací, kteří získají cenný kov v každém závodě, kterého se zúčastní (třeba Michael Phelps). Pro reprezentaci státu jako takového je důležité posílat ve výpravě sportovce, kteří mají co možná největší šanci uspět na největším svátku sportu. V rozhodování koho ještě poslat a koho už ne by mohla figurovat i umělá inteligence. Sledováním sportovců v průběhu jejich kariéry nám poskytne data, ze kterých je možné predikovat jejich potenciální úspěch na svátku všech sportů. V celém souhrnu by se tímto způsobem mohlo optimalizovat celé složení olympijské výpravy, která by měla skutečně nejlepší naděje na úspěch v podobě co možná největšího zisku medailí.

Cíl projektu

Cílem je **predikovat počet medailí reprezentačního týmu** s co nejvyšší přesností na základě dostupných údajů z předchozích sezón.

Datové soubory

K dispozici jsou různé údaje o reprezentantech účinkujících v OH, včetně jejich výsledků (informace o medailích). Informace o jednotlivcích mohou sloužit pro odvození příznaků (např. celkový počet reprezentantů v týmu apod.) pro trénování predikčního modelu. Údaje o výsledcích jednotlivců mohou sloužit pro odvození celkového počtu medailí pro daný tým a danou sezónu, který bude sloužit jako očekávaná výstupní hodnota modelu pro dané období. Jelikož se jedná o predikční úlohu, pro trénování modelu by se měla použít pouze data z období předcházejícího okamžiku predikce, tj. např. pro predikci počtu medailí v r. 2020 lze využít pouze data z předchozího období (tj. z jedné či několika posledních sezón, např. data pouze z r. 2016 či data z r. 2012 a r. 2016). Délka sledovaného a analyzovaného časového okna přitom musí respektovat známé souvislosti mezi predikovanou hodnotou a použitými příznaky (např. empirické údaje z literatury, logické uvažování).

dataOH.csv – shrnuje údaje o jednotlivých reprezentantech účinkujících v OH, včetně informací o týmu, vykonávaném sportu, výsledku, apod.

ID: ID sportovce

Name: jméno sportovce

Sex: pohlaví

Age: věk

Height: výška (cm)

Weight: váha (kg)

Team: tým

NOC: kód země NOV

Games: název OH

Year: ročník OH

Season: sezóna (Summer/Winter)

City: místo konání OH

Sport: sport

Event: sportovní událost

Medal: výsledek - medaile (NA/Bronze/Silver/Gold) (použijte pro odvození očekávaných hodnot počtu medaili pro daný stát)

POZOR:

- V datovém souboru se mohou vyskytovat odlehlé hodnoty, chybějící hodnoty měřených veličin (označené jako např. NaN, '-' apod.), hodnoty nedávající smysl apod.
- Pro vypracování projektu **používejte pouze údaje do roku 2006 (včetně)**

Výstup projektu

Výstupem projektu budou počty medaili **pro každou sledovanou sezónu (letní nebo zimní)** odvozené s využitím údajů z jedné či více předchozích sezón.

Pro vyhodnocení úspěšnosti predikce použijte *průměrnou absolutní odchylku* (mean absolute error, MAE) mezi predikovaným počtem medaili y a očekávaným/skutečným počtem d (n – počet predikovaných hodnot):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - d_i|$$

Další informace o odhadu počtu medaili v OH na základě různých parametrů např. zde:

<https://core.ac.uk/download/pdf/79557849.pdf>

<http://aassjournal.com/article-1-799-en.pdf>

KONZULTACE, TESTOVÁNÍ ALGORITMŮ A OBHAJOBA PROJEKTŮ

Konzultace

Projekty bude možné konzultovat s přiřazeným vyučujícím. Konzultace si domlouvejte individuálně přímo s vyučujícím.

Nezávislé testování odevzdaných algoritmů

Odevzdané algoritmy **budou testovány na nezávislém testovacím souboru dat** (testovací soubory mají formát shodný s trénovacími soubory). Finální nezávislé testování bude zajištěno vyučujícími. Z tohoto důvodu vás žádáme o odevzdání výstupů ve standardizovaném formátu, viz poskytnuté kódy ke každému zadání. Výstupy odevzdá vybraný člen týmu (kontaktní osoba) za celý tým. O výsledcích testování se dozvíte v rámci obhajoby projektů.

Obhajoba projektu

Řešiteli projektu budou týmy složené z **maximálně 3 studentů/studentek**.

Součástí úspěšného splnění zadání bude obhajoba projektu (prezenčně v hodině nebo MS Teams, dle aktuální situace). Prezentace se účastní všichni členové týmu. V rámci obhajoby se každý tým vyjádří ke třem částem projektu:

1. Příprava a předzpracování dat

Tato oblast může zahrnovat: parsování dat, škálování či standardizaci příznaků, ověření rovnoměrnosti zastoupení jednotlivých klasifikačních skupin a případnou augmentaci méně zastoupené skupiny dat, detekce a odstranění odlehlých hodnot, nahrazování chybějících hodnot, interpolaci, převzorkování, překódování vstupních/výstupních dat, způsob rozdělení na trénovací a testovací množinu, selekce vhodných příznaků,...)

1. Použitá predikční/klasifikační metoda

Např.: Výběr vhodného modelu, nastavení modelu, ověření úspěšnosti predikce/klasifikace,...

1. Interpretace a diskuze výsledků

Např.: Rozebrání použitých příznaků a parametrů modelu a jejich vlivu na výsledek, hodnocení generalizační schopnosti modelu, interpretace výsledků vzhledem k povaze zadaného problému, porovnání s výsledky z literatury, limitace a výhody použitého postupu...)

Přiřazení studenta k části prezentace bude provedeno náhodně, proto každý student musí být připraven prezentovat jakoukoli část projektu. Hodnocená bude odborná úroveň projektu i schopnost reagovat na případné dotazy publika a komise.

Plagiátorství bude mít za následek neudělení zápočtu!

Délka obhajoby projektu je max 12 min, včetně prezentace výsledků a diskuze. Tomu odpovídá cca 5-8 slajdů prezentace (*.ppt) a 5 min diskuze. Účast týmů na prezentaci výsledků ostatních skupin není povinná. Aktuální informace o obhajobě projektů bude průběžně zveřejňována v eLearningu.

DŮLEŽITÉ TERMÍNY

Zaslání složení týmů, vybraného tématu projektu a vybraného prostředí pro vypracování zadání projektu (Matlab, Python) na ronzhina@vut.cz (informace za celý tým zasílá kontaktní osoba): **22. října**

Nejzazší termín odevzdání řešení (stačí kontaktní osoba): **8. prosince**

Termín obhajoby: **14. prosince** v rámci výukových hodin (dle rozvrhu). Pořadí prezentací bude předem dohodnuto s týmy.