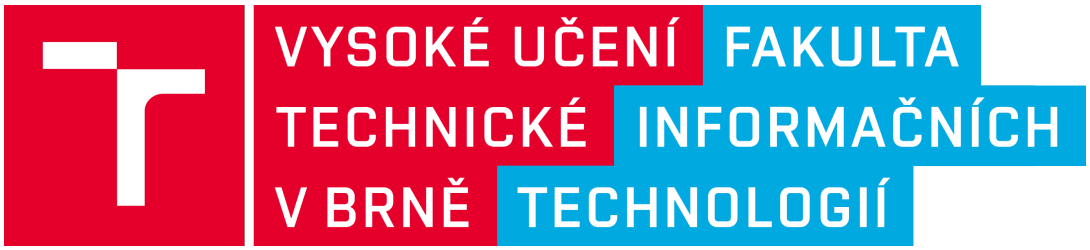


ZZN Projekt

Získávání znalostí z databáze diabetiků

(Řešení)

Bc. Ján Maťufka (xmatuf00)

Bc. Daniel Prudký (xprudk08)

2024

Popis datasetu

Dataset „Diabetes 130-US hospitals for years 1999-2008“ obsahuje informace o hospitalizacích více než 100 000 pacientů s diabetem ze 130 různých nemocnic ve Spojených státech během let 1999 až 2008. Obsahuje demografická data pacientů, jako je věk, pohlaví, rasa, a informace o jejich diagnózách, laboratorních výsledcích a medikaci, včetně léčby inzulínem. Kromě toho zahrnuje údaje o hospitalizačních výstupech, jako je opakovaná hospitalizace a délka pobytu. Detailnější popis datasetu a popis jednotlivých sloupců je dostupný na webu[1].

Formulace úloh

V našem projektu budeme řešit čtyři úlohy získávání znalostí z tohoto datasetu.

1: Predikce opakované hospitalizace

V této úloze budeme chtít umět klasifikovat pacienty z hlediska rehospitalizace. Na základě testů a dalších charakteristik umět predikovat, jestli pacient bude znovu hospitalizován do 30 dnů, po více než 30 dnech, nebo pacient opakovanou hospitalizaci nebude potřebovat.

2: Analýza korelací mezi typem léčby a výsledkem hospitalizace

Cílem této úlohy je analyzovat vztahy mezi různými typy léčby diabetu a výsledky hospitalizace, jako jsou opakovaná hospitalizace, komplikace, či úspěšná stabilizace pacienta. Pomocí této analýzy chceme zjistit, které léčebné postupy vedou k pozitivním výsledkům a které mohou být spojeny s horšími výstupy.

3: Predikce délky hospitalizace

Cílem této úlohy je předpovědět délku hospitalizace pacienta na základě jeho demografických údajů, diagnóz a typu léčby, což by mohlo výrazně pomoci při efektivnějším plánování zdravotnických zdrojů a péče. Zároveň by výsledky měli pomoci lépe odhadovat nároky na hospitalizaci pro jednotlivé pacienty.

4: Analýza dopadu demografických faktorů na výsledky hospitalizace

Na rozdíl od první úlohy je tato úloha zaměřena na výsledek prvotní hospitalizace a taky se tu bude větší důraz klást na faktory jako je věk, pohlaví, rasa, než na medicínske informace a výsledky různých testů. Zajímá nas tedy, jak diabetes působí na různé části populace, případně které demografické skupiny jsou nejvíc rizikové.

[1] `https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008`

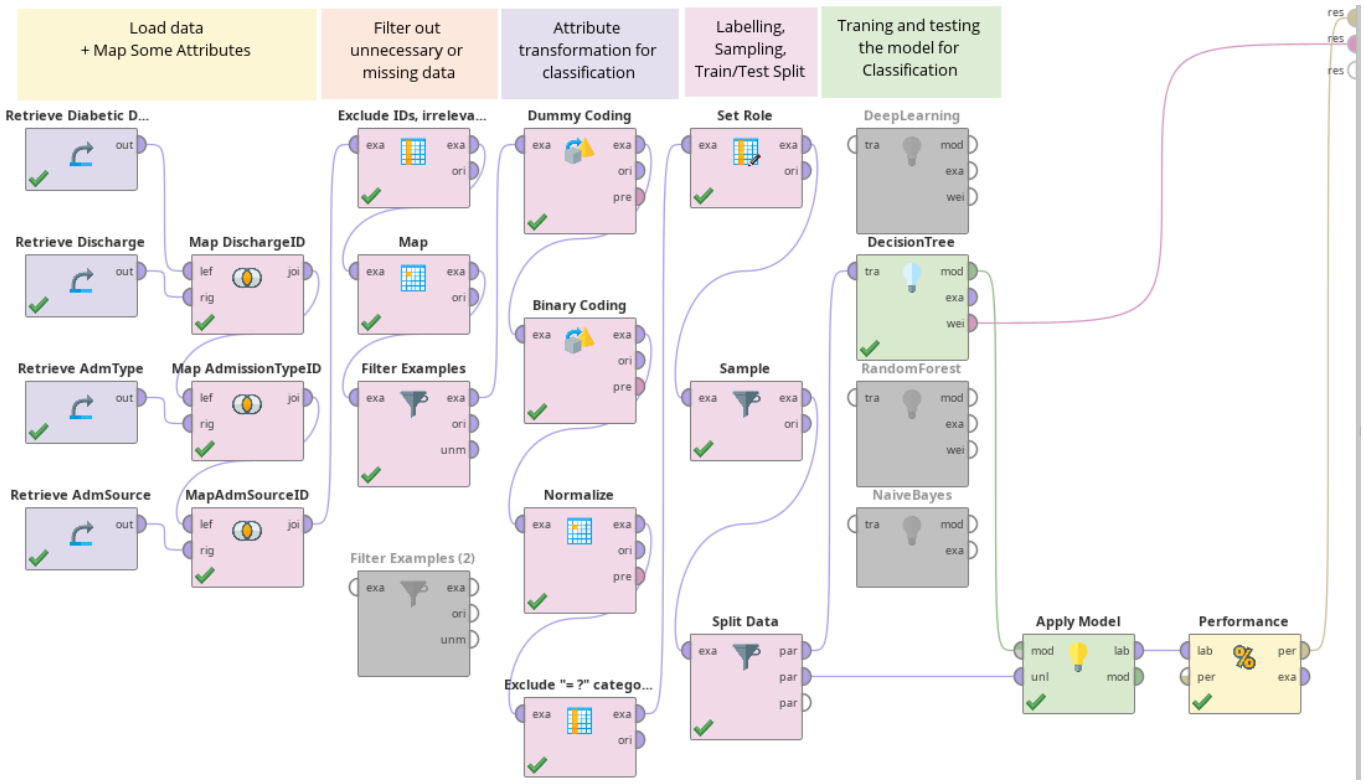
Řešení úlohy 1: Predikce opakované hospitalizace

V této úloze jsme se snažili klasifikovat pacienty na 3 třídy (atribut `readmitted`):

- `NO`: Pacienti, kteří posléze nebyli hospitalizování.
- `>30`: Pacienti, kteří byli opakovaně hospitalizování po více než 30 dnech.
- `<30`: Pacienti, kteří byli opakovaně hospitalizování dřív než po 30 dnech.

Pro klasifikaci jsme chtěli porovnat více modelu. Různé modely mají pro správné naučení klasifikace na data různé požadavky: vyváženost tříd, normalizované numerické atributy, kategorické atributy převedeny na binární (one-hot encoding), úplnost záznamů, atd. Proto se část procesu týkající se úpravy dat skládala z částí (viz Obrázek 1):

1. *Join* – převedení identifikátorů některých atributů na jejich textový popis (byl součástí datasetu v souboru `ID_mapping.csv`). Jednalo se o atributy:
   - `discharge_disposition_id`,
   - `admission_source_id`,
   - `admission_type_id`.
2. *Select Attributes* – odmazání zbytečných atributů nebo atributů s řídkym výskytem:
   - identifikátory pacienta a návštěvy,
   - již přemapovány identifikátory z předešlého kroku,
   - kód platby (pojišťovna), specializace vyšetřujicího doktora,
   - váha pacienta (jen 3 % záznamů obsahují tuhle informaci).
3. *Filter Examples* – odfiltrování záznamů z chybějicími hodnotami (předtím se ještě sjednotili řetězce reprezentující chybějící záznam pomocí *Map*).
4. Transformace atributů:
   - *Nominal to Numerical* – překódování kategorických atributů do binárních.
   - *Normalize* – min-max normalizace kvantitativních atributů.
5. Příprava pro klasifikaci:
   - *Set Role* – nastavení atributu pro klasifikaci (`readmitted` nastaven jako `label`).
   - *Sample* – rovnoměrné navzorkování instancí jednotlivých tříd. Původný dataset obsahuje přes 101 tisíc záznamů: 54 % `NO`, 35 % `>30`, a jen 11 % záznamů patří do třídy `<30`. Po odfiltrování neúplných záznamů a po navzorkování měly všechny třídy zastoupení 9 000 záznamů. Celkem se tedy pro klasifikování (trénovaní a testování) použilo 27 000 záznamů.
   - *Split Data* – rozdělení na trénovací data (80 % – 21 600 záznamů) a testovací data (20 % – 5 400 záznamů).



Obrázek 1: Schéma procesu pro úlohu 1 (predikce opakované hospitalizace)

Pro klasifikaci jsme vybrali 4 modely. Níže jsou vyjmenovány, spolu s podstatnými parametry, na základě kterých byly dané modely natrénovány:

1. **Hluboké učení** – *Deep Learning* – ponechány výchozí atributy.
2. **Rozhodovací strom** – *Decision Tree*
   - rozhodovací kritérium: gini index,
   - maximální hloubka stromu: 9,
   - prořezávaní (pruning): ano,
   - pravděpodobnostní práh pro prořezávání (pruning confidence): 0.1,
   - předprořezávaní (prepruning): ne,
   - ostatní atributy ponechány s výchozími hodnotami.
3. **Náhodný les** – *Random Forest*
   - rozhodovací kritérium: gini index,
   - maximální hloubka stromu: 12,
   - prořezávaní (pruning): ano.
   - pravděpodobnostní práh pro prořezávání (pruning confidence): 0.1
   - ostatní atributy ponechány s výchozími hodnotami.
4. **Naivní Bayesův klasifikátor** – *Naive Bayes* – ponechány výchozí atributy.
   - Poznámka: Při tomto modelu bylo nutné odstránit také atributy o diagnózách pacientů (`diag_1`, `diag_2`, `diag_3`). Jelikož se jedná o kategorické atributy, které můžou nabývat stovky různých hodnot, jejich transformace by značně zvětšila dataset a spomalila proces učení. Ponechání diagnóz v původním stavu taky vedlo k méně přesnému modelu (značně preferoval jednu třídu).

Výsledky celkové přesnosti predikce jednotlivých modelů:

Deep Learning – 45.11 %

- přesnosti predikce pro jednotlivé třídy – viz Tabulka 1
- váhy významností jednotlivých atributů – viz Tabulka 2

Decision Tree – 43.31 %

- přesnosti predikce pro jednotlivé třídy – viz Tabulka 3
- váhy významností jednotlivých atributů – viz Tabulka 4

Random Forest – 46.22 %

- přesnosti predikce pro jednotlivé třídy – viz Tabulka 5
- váhy významností jednotlivých atributů – viz Tabulka 6

Naive Bayes – 38.06 %

- přesnosti predikce pro jednotlivé třídy – viz Tabulka 7

Celkově se ukázal jako nejlepší model pro predikci náhodný les. Při modelech, které umožňují nahlédnout do významnosti atributů využívaných při predikci (Tabulky 2, 4, 6), se ukazují tyto faktory jako nejvíce významné:

- `age` – věk pacienta,
- `num_medications` – počet léků, které pacient užívá,
- `num_lab_procedures` – počet různých testů/odběrů, které byly na pacientovi vykonány,
- `time_in_hospital` – délka hospitalizace pacienta (v dnech).

Přesnost klasifikace všemi použitými modely se pohybuje kolem 38–46 %, co je jen marginálně lepší, než náhodně tipovat. To může znamenat, že na základě poskytnutých dat není možné efektivně predikovat, jestli pacient bude rehospitalizován, nebo nikoliv. To dáva smysl, protože data obsahují málo informací, většina z nich se týká užívaných léků. Data neobsahují informace o symptomech, životních podmínkách, stravě pacientů, nebo hlubší zdravotní historii pacientů. Tyto informace doktoři určite při diagnóze zohledňují, a jistě by vedly na mnohem lepší predikci.

Tabulka 1: Přesnost predikce úlohy 1 hlubokým učením

accuracy: 45.11%

	true NO	true >30	true <30	class precision
pred. NO	1129	716	586	46.44%
pred. >30	283	478	385	41.71%
pred. <30	388	606	829	45.47%
class recall	62.72%	26.56%	46.06%	

Tabulka 2: Soupis nejvýznamnějších atributů pro predikci hlubokým učením s jejich váhami významností

attribute	weight ↓
number_inpatient	1
discharge_disposition = Expired	0.724
change	0.701
gender	0.645
discharge_disposition = Discharged to home	0.620
time_in_hospital	0.607
number_diagnoses	0.597
number_emergency	0.596
num_procedures	0.587
admission_type = Emergency	0.587
discharge_disposition = Discharged/transferred to another rehab f...	0.584
age	0.580
number_outpatient	0.578
admission_source = Physician Referral	0.565
insulin = Steady	0.549
admission_source = Emergency Room	0.547
num_medications	0.536
diag_3.403	0.525
diabetesMed	0.522
num_lab_procedures	0.515
race = Caucasian	0.497
admission_type = Urgent	0.492
diag_1.428	0.482
diag_1.V58	0.467
diag_3.250	0.451

Tabulka 3: Přesnost predikce úlohy 1 rozhodovacím stromem

accuracy: 43.31%

	true NO	true >30	true <30	class precision
pred. NO	921	588	448	47.06%
pred. >30	521	732	666	38.14%
pred. <30	358	480	686	45.01%
class recall	51.17%	40.67%	38.11%	

Tabulka 4: Soupis nejvýznamnějších atributů pro predikci rozhodovacím stromem s jejich váhami významností

attribute	weight ↓
age	0.249
num_medications	0.114
time_in_hospital	0.104
num_lab_procedures	0.097
A1Cresult = >7	0.061
num_procedures	0.057
admission_type = Urgent	0.044
gender	0.042
diabetesMed	0.041
glipizide = No	0.027
glyburide = Steady	0.024
race = AfricanAmerican	0.020
discharge_disposition = Discharged/transferred to home with home health service	0.017
A1Cresult = Norm	0.013
number_outpatient	0.012
number_diagnoses	0.011
admission_source = Emergency Room	0.009
insulin = Steady	0.007
number_emergency	0.006
number_inpatient	0.006
discharge_disposition = Discharged/transferred to another short term hospital	0.006
glimepiride = Up	0.005
repaglinide = No	0.005
admission_source = Transfer from another health care facility	0.004
insulin = Down	0.003

Tabulka 5: Přesnost predikce úlohy 1 náhodným lesem

accuracy: 46.22%

	true NO	true >30	true <30	class precision
pred. NO	1086	665	521	47.80%
pred. >30	312	556	425	43.00%
pred. <30	402	579	854	46.54%
class recall	60.33%	30.89%	47.44%	

Tabulka 6: Soupis nejvýznamnějších atributů pro predikci náhodným lesem s jejich váhami významností

attribute	weight ↓
num_lab_procedures	0.098
num_medications	0.086
time_in_hospital	0.082
age	0.070
num_procedures	0.054
number_inpatient	0.050
number_diagnoses	0.047
number_outpatient	0.030
gender	0.029
number_emergency	0.020
admission_type = Urgent	0.019
race = Caucasian	0.018
admission_type = Emergency	0.017
metformin = No	0.017
change	0.017
insulin = No	0.016
insulin = Steady	0.015
metformin = Steady	0.014
admission_type = Elective	0.013
race = AfricanAmerican	0.013
admission_source = Physician Referral	0.013
admission_source = Emergency Room	0.013
A1Cresult = None	0.013
glipizide = No	0.012
insulin = Up	0.012

Tabulka 7: Přesnost predikce úlohy 1 naivním Bayesovým klasifikátorem

accuracy: 38.06%

	true NO	true >30	true <30	class precision
pred. NO	96	20	32	64.86%
pred. >30	1403	1478	1287	35.46%
pred. <30	301	302	481	44.37%
class recall	5.33%	82.11%	26.72%	

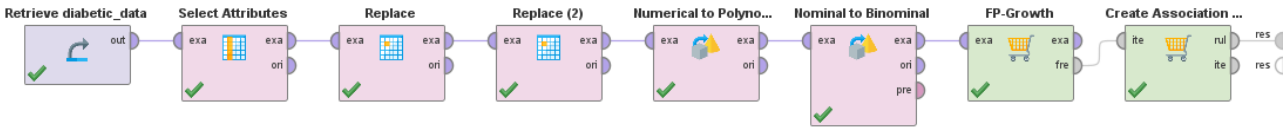
Řešení úlohy 2: Analýza korelací mezi typem léčby a výsledkem hospitalizace

K analýze korelací jsme použili asociační pravidla, s nimiž RapidMiner nabízí jednoduchou práci. K analýze jsme vyvužili sloupce obsahující záznamy dávkování léků a sloupec `discharge_disposition_id`, který dokumentuje dispozice k propuštění pacienta. Ve výsledcích se nám pak výrazněji projevili následující id:

- `id = 1` – Propuštěn domů
- `id = 3` – Propuštění/přeložení do sociálního zařízení
- `id = 6` – Propuštění/převedení domů s domácí zdravotní službou

Při modelování pipeline jsme nejprve připravili data, tedy vybrali atributy a převedli na binominální hodnoty. K převodu záznamů o lécích jsme využili blok *Replace*. Záznamy v těchto sloupcích udávají hodnoty [*No, Steady, Up, Down*], proto jsme je zgeneralizovali tak, že jsme hodnoty *No* nastavili na *false*, jelikož znamenají, že pacient nebere dané léky a ostatní na *true*, protože to znamená, že pacient dané léky užíval. Pro discharge_disposition_id jsme pomocí bloků *Numerical to Polynominal* a *Nominal to Binominal* vytvořili dummy variables přiřadili jim binominální hodnotu na základě přítomnosti v dané třídě.

Při vytváření asociačních pravidel jsme použili bloky *FP-Growth* a *Create Association Rules*. Nejprve jsme si vytvořili frekvenční množiny s omezeními minimální podpory 0.02, intervalem velikosti množin $\langle 2, 5\rangle$ a s parametrem `must contain regexp = "discharge_disposition_id_.*"`, aby jsme dostávali pouze pravidla spojená s výsledkem léčby. Minimální podpora 0.02 je velice nízká, jen málo výsledných pravidel ale bohužel dosahovalo vyšších podpor, což naznačuje že korelace mezi užívanými léky a testovanou dispozicí propouštění nebude moc silná. Z frekvenčních množil jsme následně vytvořili asociační pravidla, která jsme také filtrovali minimální spolehlivostí 0.3. Celá pipeline je na Obrázku 2.



Obrázek 2: Design řešení úlohy 2

Výsledky analýzy jsou v Tabulce 8. Co se týče dispozice *Propuštěn domů* tak největší podporu má společně s inzulínem, metforminem, glipizidem, kombinací inzulínu a metforminu a nad 5 % dosahuje ještě glyburid. Další dvě výše zmíněné dispozice s id 3 a 6 dosahují vyšší podpory pouze s inzulínem, konkrétně zhruba 7–8 %, a metforminem (2–3 %).

Všem dvouprvkovým frekvenčním množinám vypočítal RapidMiner spolehlivost ∞, což by mohlo znamenat, že podpora premisy je téměř nulová a dochází k dělení nulou nebo je podpora průniku premisy

a závěru téměř nulová a tedy by se spolehlivost také blížila k nule. Pak by RapidMiner vypsal ∞. Po ručním výpočtu na příkladu frekvenční množiny obsahující `insulin` a `discharge_disposition_id_1`, kdy $supp(\texttt{insulin}) = 0.534$ a $supp(\texttt{insulin} \cap \texttt{discharge\_disposition\_id\_1}) = 0.316$, pak by měla být spolehlivost zhruba 0.592. Tuto chybu výpočtu se mi nepodařilo vyřešit.

Mezi víceprvkovými frekvenčními množinami analýza dosahuje spolehliosti 57-63%. Ostatní dispozice k propuštění pacienta nedosahovali ani velice nízko nastavených minim, především minuma podpory při tvorbě frekvenčních množin, a proto nejsou zmíněny. Když jsme minimální podporu snížili na 0.1 ve výsledcích se objevili i id 2 a 22, tedy Propuštění/přeložení do jiné krátkodobé nemocnice a propuštění/přeložení na jinou rehabilitační fakultu včetně rehabilitačních oddělení nemocnice.

Tabulka 8: Výpis nejvýznamnějších asosiačních pravidel pro úlohu 2

Premises	Conclusion ↑	Support	Confiden... ↓	LaPlace	Gain	p-s	Lift	Convicti...
insulin	discharge_disposition_id_1	0.316	∞	1.316	0.316	0.316	∞	-0
metformin	discharge_disposition_id_1	0.123	∞	1.123	0.123	0.123	∞	-0
glipizide	discharge_disposition_id_1	0.074	∞	1.074	0.074	0.074	∞	-0
glyburide	discharge_disposition_id_1	0.058	∞	1.058	0.058	0.058	∞	-0
pioglitazone	discharge_disposition_id_1	0.044	∞	1.044	0.044	0.044	∞	-0
rosiglitazone	discharge_disposition_id_1	0.038	∞	1.038	0.038	0.038	∞	-0
glimepiride	discharge_disposition_id_1	0.032	∞	1.032	0.032	0.032	∞	-0
metformin, glipizide	discharge_disposition_id_1	0.022	0.633	0.988	-0.048	0.022	∞	2.724
insulin, metformin	discharge_disposition_id_1	0.061	0.623	0.966	-0.136	0.061	∞	2.649
insulin, pioglitazone	discharge_disposition_id_1	0.024	0.600	0.985	-0.055	0.024	∞	2.503
insulin, rosiglitazone	discharge_disposition_id_1	0.020	0.600	0.987	-0.048	0.020	∞	2.502
metformin, glyburide	discharge_disposition_id_1	0.022	0.582	0.985	-0.053	0.022	∞	2.395
insulin, glipizide	discharge_disposition_id_1	0.035	0.582	0.976	-0.086	0.035	∞	2.393
insulin, glyburide	discharge_disposition_id_1	0.025	0.572	0.982	-0.063	0.025	∞	2.338

Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Convicti...
insulin	discharge_disposition_id_6	0.069	∞	1.069	0.069	0.069	∞	-0
metformin	discharge_disposition_id_6	0.028	∞	1.028	0.028	0.028	∞	-0

Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Convicti...
insulin	discharge_disposition_id_3	0.082	∞	1.082	0.082	0.082	∞	-0
metformin	discharge_disposition_id_3	0.022	∞	1.022	0.022	0.022	∞	-0

Řešení úlohy 3: Predikce délky hospitalizace

Predikce atributu `time_in_hospital` na základě aktributů dostupných při hospitalizaci pacienta. Pro predikci byli zvoleny veškeré atributy, které popisují demografické údaje a údaje dostupné při hospitalizaci pacienta. Nebyli vybrány léky, atributy popisující výsledný stav pacienta ani jeho váha (kvůli příliž malému zastoupení). Zvoleny tedy byli následovné:

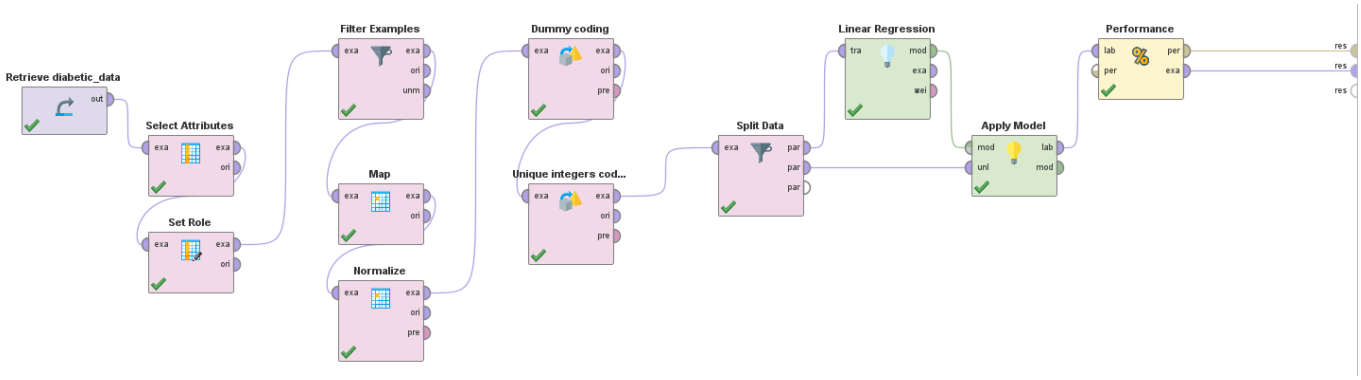
- podmínky přijetí pacienta: `admission_source_id`, `admission_type_id`,
- demografické údaje: `age`, `gender`, `race`,
- diagnózy: `diag_1`, `diag_2`, `diag_3`,
- statistiky o předešlých hospitalizacích: `number_diagnoses`, `number_emergency`, `number_inpatient`, `number_outpatient`,
- `medical_specialty`, `payer_code`.

.

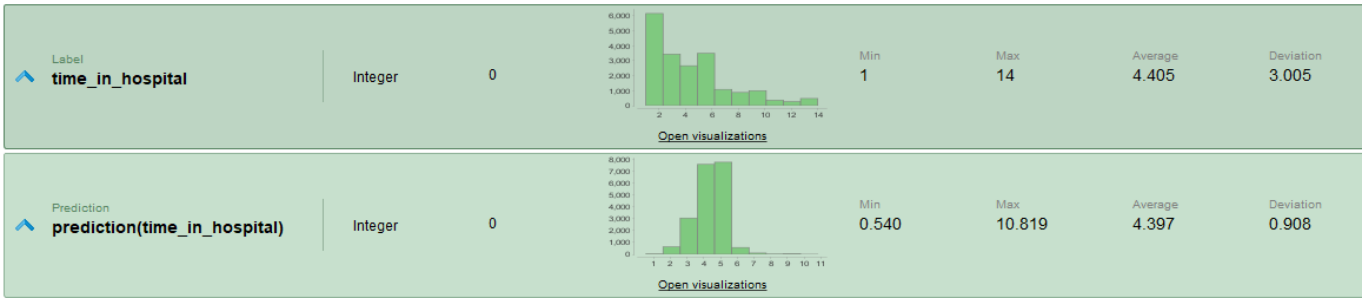
Při přípravě dat jsme využili několik různých bloků. Prvním byl *Select Attributes*, ve kterém jsme si extrahovali zvolené sloupce. Druhý blok *Set Role* určuje sloupec, jehož hodnotu se snažíme predikovat. Následují bloky *Filter Examples, Map a Normalize*, které zajišťují čištění dat od prázdných hodnot, mapují věk z intervalů na průměrný daného intervalu a normalizují numerické hodnoty. V blocích *Nominal to numerical* nazvaných podle příslušného typu kódování převádíme kategorická data na numerická pomocí dummy a unique values kódování. V posledním, *Split Data*, bloku rozdělujeme data na trénovací(80%) a testovací část(20%). Data rozdělujeme pomocí stratified sampling, což nám zajišťuje, že cílové proměnná je zastoupena v tréninkových i testovacích sadách v poměru k jejímu výskytu v celém datasetu.

Pro vytvoření modelu jsme použili Lineární regresy. Vyzkoušeli jsme všechny možnosti feature selection, které RapidMiner nabýzí, ale žádná nepřinasla lepší výsledky. Po aplikování modelu a výpočtu performance jsme byli schopni natrénovat model s *Root Mean Squared Error* 2.858.

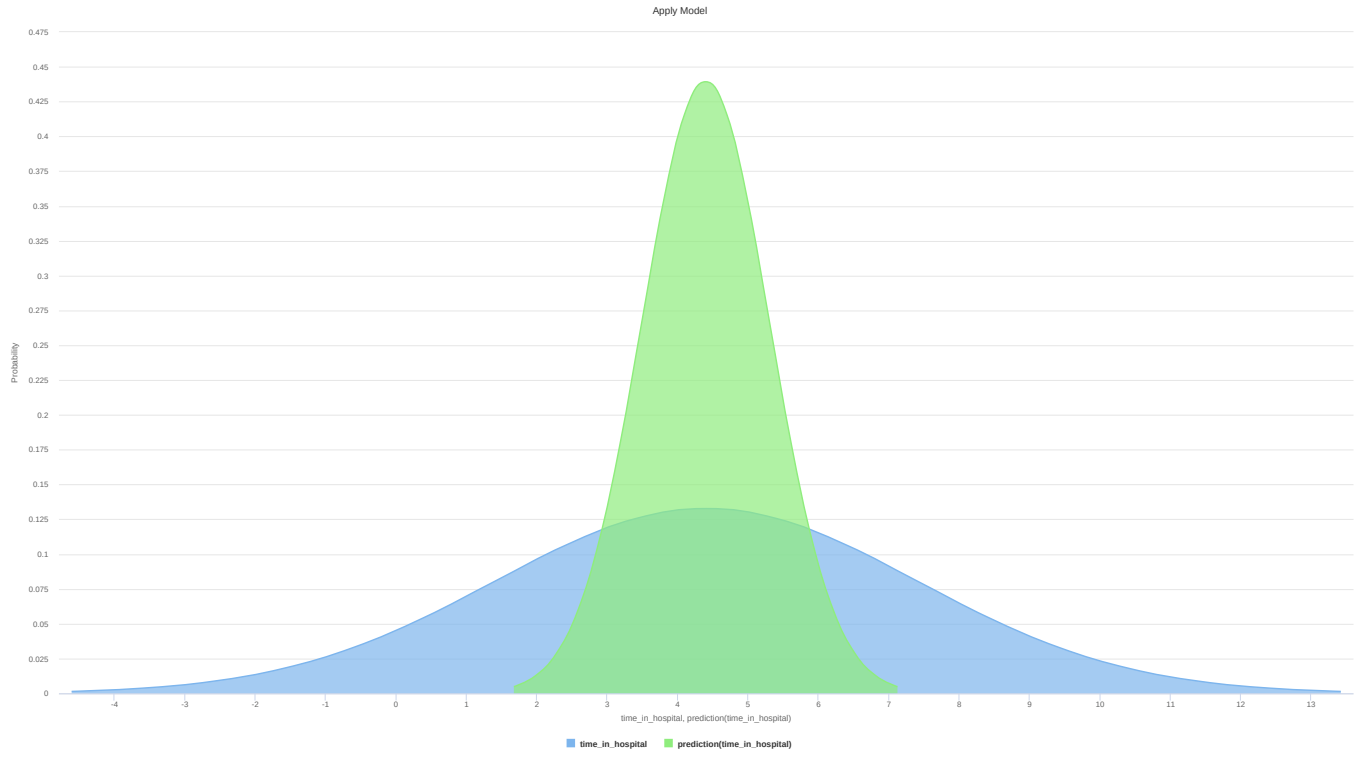
Výsledky poukazují na omezenou schopnost odhadnout délku hospitalizace pacienta v momentě příjmu, pouze s existujícími znalostmi.



Obrázek 3: Design řešení úlohy 3



Obrázek 4: Statistiky predikce úlohy 3



Obrázek 5: Graf predikce z úlohy 3 v porovnání s původní hodnotou

Řešení úlohy 4: Analýza dopadu demografických faktorů na výsledky hospitalizace

Tato úloha se snažila zjistit vazby mezi demografickými faktory (věk, pohlaví, rasa) a výsledkem hospitalizace (atribut `discharge_disposition_id`). Tuto analýzu jsme udělali 3 způsoby:

1. **Explorativní analýza dat** – sledování sloupcových grafů závislostí výsledku hospitalizace od jednotlivých faktorů.
2. **Sestavení prediktivního modelu** – na základě váh významností daných atributů na přesnost predikce můžeme usuzovat, které faktory mají jak silný dopad.
3. **Korelační analýza** – test dobré shody dvojic atributů.

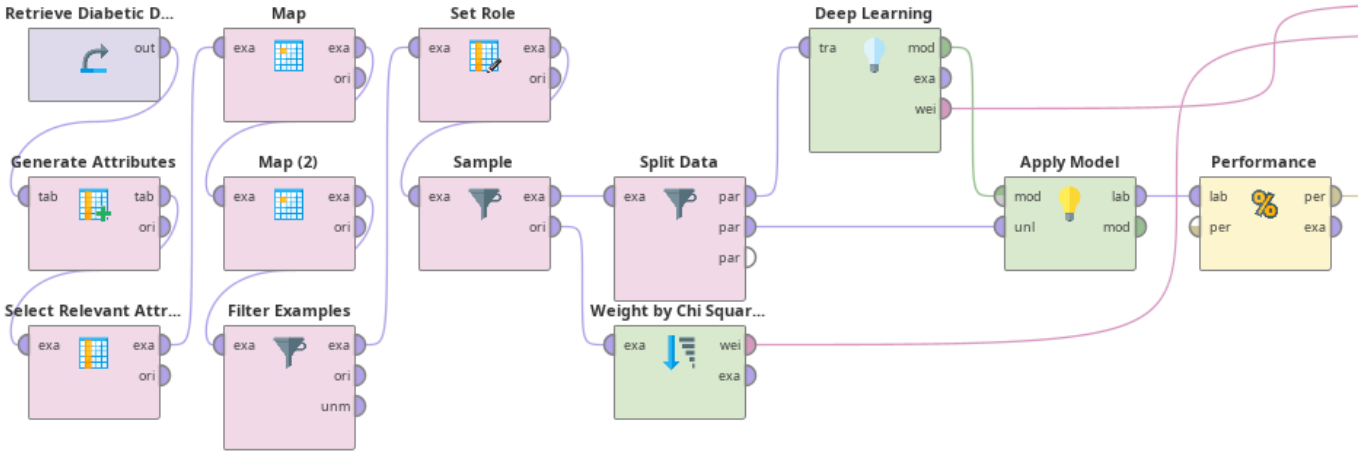
Atribut `discharge_disposition_id` může nabývat 30 různych hodnot. Protože sledování tolik různých výsledků hospitalizace by bylo nepřehledné, byli tyto hodnoty združeny do 5 kategórií: Pozitivní výsledek hospitalizace, neutrální výsledek, negativní výsledek, neznámy výsledek/chybějící hodnota, pořád v péči (ještě neznámy výsledek). Způsob mapování je znázorněn na Tabulce 9.

Tabulka 9: Tabulka mapování různých výsledků hospitalizace pro sjednodušení úlohy

ID	Description	Category
1	Discharged to home	Positive
2	Discharged/transferred to another short term hospital	Neutral
3	Discharged/transferred to SNF	Neutral
4	Discharged/transferred to ICF	Neutral
5	Discharged/transferred to another type of inpatient care institution	Neutral
6	Discharged/transferred to home with home health service	Positive
7	Left AMA	Unknown
8	Discharged/transferred to home under care of Home IV provider	Positive
9	Admitted as an inpatient to this hospital	Still Under Care
10	Neonate discharged to another hospital for neonatal aftercare	Still Under Care
11	Expired	Negative
12	Still patient or expected to return for outpatient services	Still Under Care
13	Hospice / home	Negative
14	Hospice / medical facility	Negative
15	Discharged/transferred within this institution to Medicare approved swing bed	Neutral
16	Discharged/transferred/referred another institution for outpatient services	Neutral
17	Discharged/transferred/referred to this institution for outpatient services	Neutral
18	NULL	Unknown
19	Expired at home. Medicaid only, hospice.	Negative
20	Expired in a medical facility. Medicaid only, hospice.	Negative
21	Expired, place unknown. Medicaid only, hospice.	Negative
22	Discharged/transferred to another rehab fac including rehab units of a hospital.	Neutral
23	Discharged/transferred to a long term care hospital.	Negative
24	Discharged/transferred to a nursing facility certified under Medicaid.	Neutral
25	Not Mapped	Unknown
26	Unknown/Invalid	Unknown
27	Discharged/transferred to a federal health care facility.	Neutral
28	Discharged/transferred/referred to a psychiatric hospital/unit.	Still Under Care
29	Discharged/transferred to a Critical Access Hospital (CAH).	Negative
30	Discharged/transferred to another Type of Health Care Institution.	Neutral

Schéma procesu znázorňující úpravu dat a získaní výsledku je na Obrázku 6. Příprava dat probíhala v následujících krocích:

1. *Generate Attributes* – vytvoření atributu `outcome`, který je jenom řetězcová (a tedy kategorická) reprezentace identifikátoru `discharge_disposition_id`.
2. *Select Attributes* – omezení datasetu na 4 atributy (`race`, `age`, `gender`, `outcome`)
3. *Map* – přemapování atributu `outcome` podle Tabulky 9, sjednocení reprezentantů neznámych hodnot.
4. Příprava pro klasifikaci – vymazání záznamů s chybějícími hodnotami (*Filter Examples*), nastavení klasifikovaného atributu (*Set Role*), rovnoměrné navzorkování datasetu – 3 třídy po 2790 záznamech (*Sample*).



Obrázek 6: Schéma procesu pro analýzu dopadu demografických dat na výsledek hospitalizace

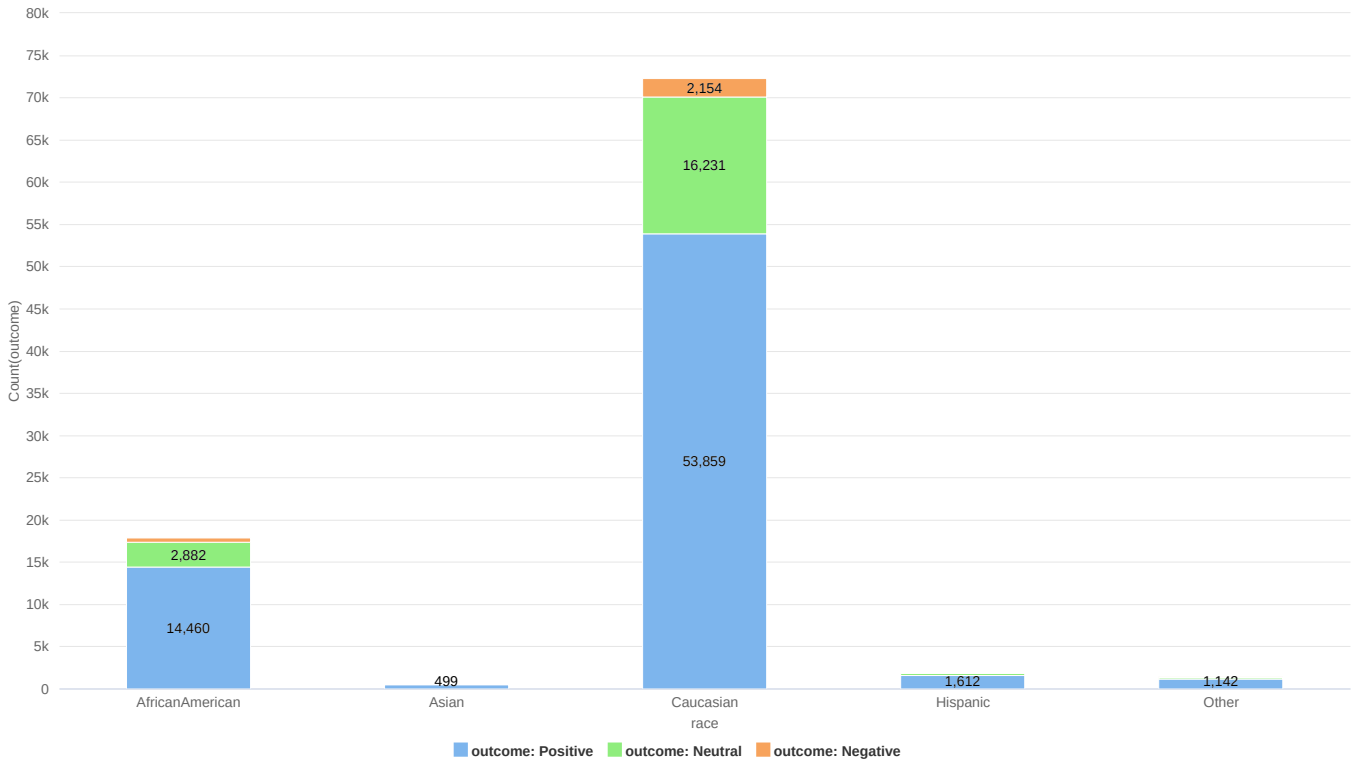
Explorativní analýza dat

Tyto výsledky byly zaznamenány po kroku *Set Role* (tedy převedením výstupu `exa` na port `res` – výstupní port procesu). Z počátečních analýz je možné udělat tyto pozorování (viz Tabulka 10, Obrázky 7, 8, 9, 10, 11, 12):

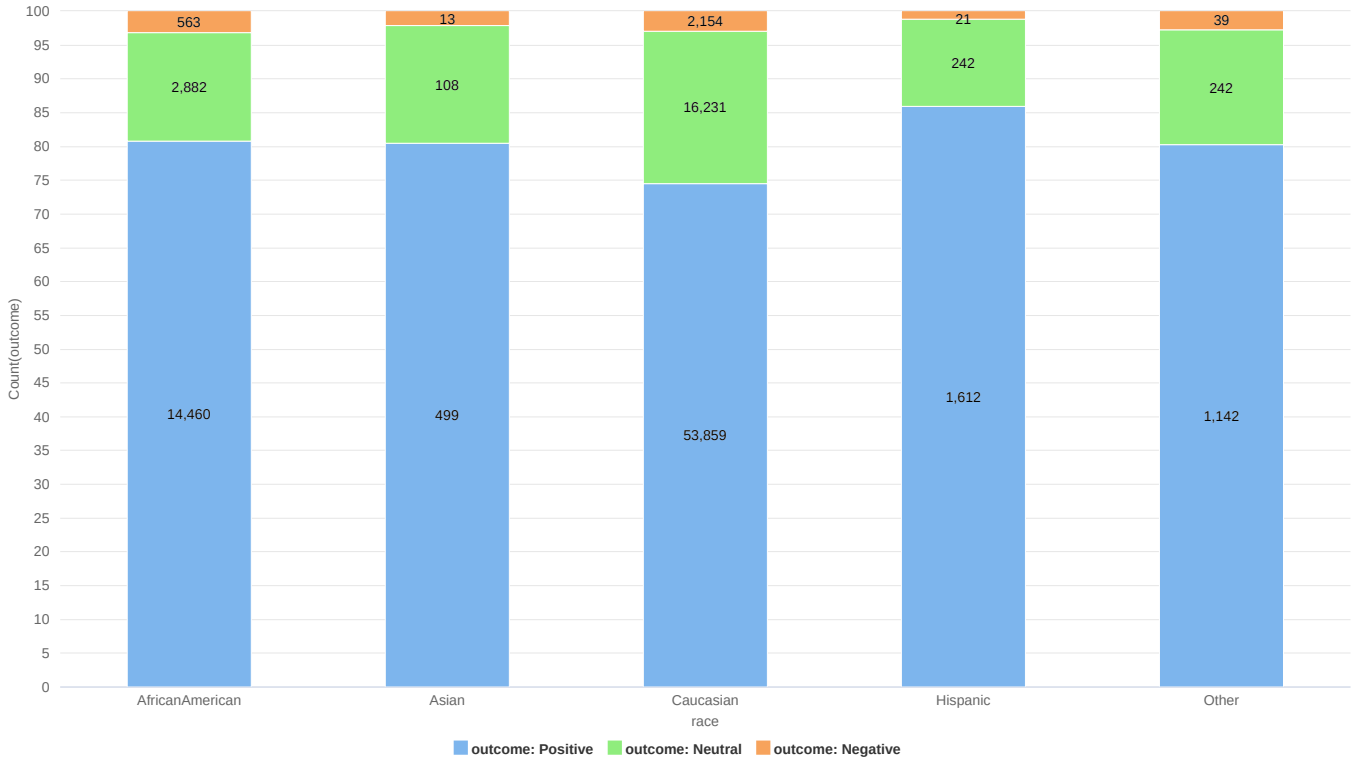
- ženy jsou o trochu více rizikovější skupinou než muži,
- nejvíce ohroženou skupinou jsou lidi nad 80 let,
- na rase z hlediska výsledků hospitalizace význačne nezáleží, vyšetření lidi bílé pleti dopadne okrajově hůř, než u ostatních ras (avšak ostatní rasy nemají zdaleka takovou četnost v rámci datasetu).

Tabulka 10: Počty a podíly hodnot jednotlivých atributů skoumaných v této analýze vůči dopadu hospitalizace. Poznámka*: První tři sloupce s procenty značí poměry výsledků hospitalizace pro hodnotu daného atributu, poslední sloupec značí poměr dané části populace v rámci celého datasetu.

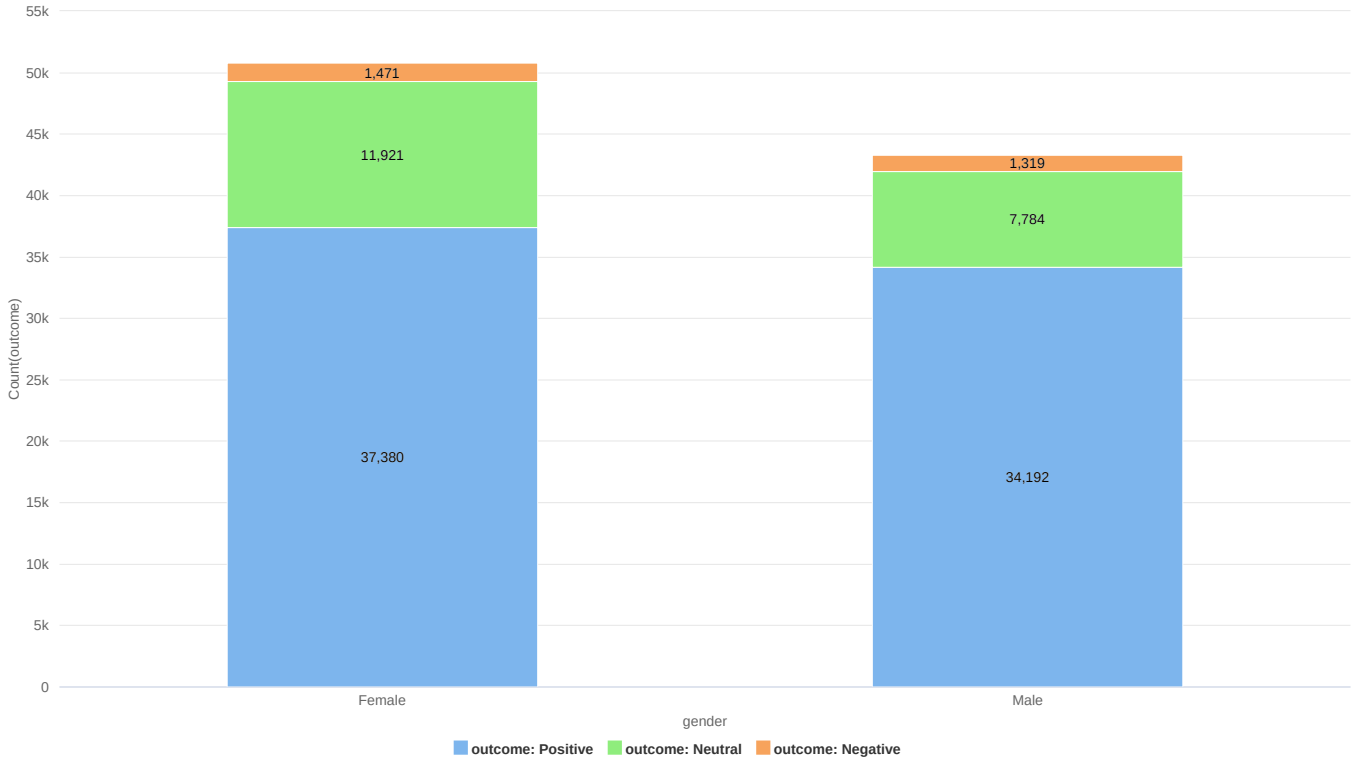
	value	positive	%	neutral	%	negative	%	sum	%
race	african-american	14 460	80.8	2 882	16.1	563	3.1	17 905	19.0
	asian	499	80.5	108	17.4	13	2.1	620	0.7
	caucasian	53 859	74.6	16 231	22.5	2 154	3.0	72 244	76.8
	hispanic	1 612	86.0	242	12.9	21	1.1	1 875	2.0
	other	1 142	80.3	242	17.0	39	2.7	1 423	1.5
gender	female	37 380	73.6	11 921	23.5	1 471	2.9	50 772	54.0
	male	34 192	79.0	7 784	18.0	1 319	3.0	43 295	46.0
age	0-10	146	98.0	2	1.3	1	0.7	149	0.2
	10-20	632	98.4	10	1.6	0	0.0	642	0.7
	20-30	1 446	96.2	46	3.1	11	0.7	1 503	1.6
	30-40	3 259	94.4	182	5.3	11	0.3	3 452	3.7
	40-50	8 091	91.7	626	7.1	103	1.2	8 820	9.4
	50-60	13 812	87.0	1 818	11.5	243	1.5	15 873	16.9
	60-70	16 961	81.6	3 316	16.0	512	2.5	20 789	22.1
	70-80	17 148	71.2	6 091	25.3	830	3.4	24 069	25.6
	80-90	9 011	55.8	6 276	38.9	853	5.3	16 140	17.2
	90+	1 066	40.5	1 338	50.9	226	8.6	2 630	2.8
sum		71 572	76.1	19 705	20.9	2 790	3.0	94 067	100.0



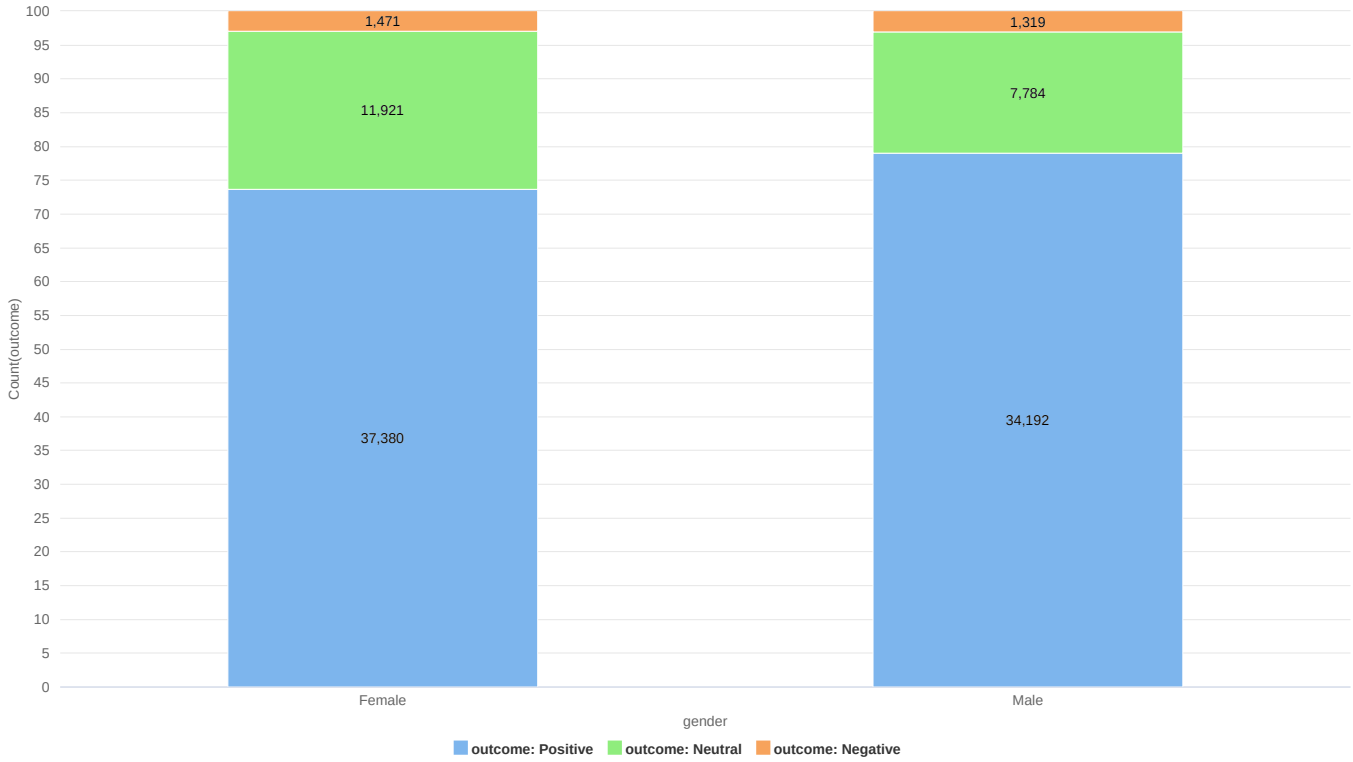
Obrázek 7: Porovnání počtu jednotlivých výsledku hospitalizace v rámci pacientů dané rasy



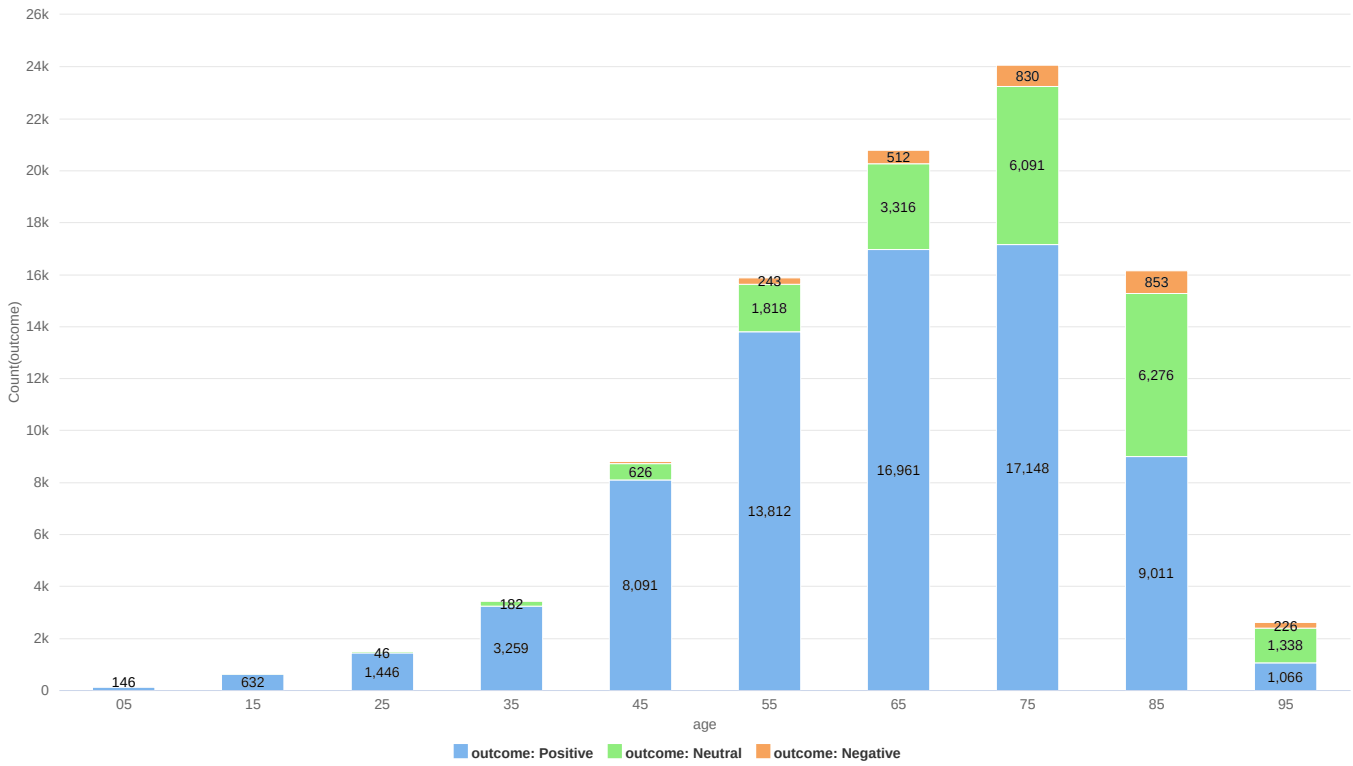
Obrázek 8: Porovnání poměrů jednotlivých výsledku hospitalizace v rámci pacientů dané rasy (y-osa značí procenta z celkového počtu daných pacientů)



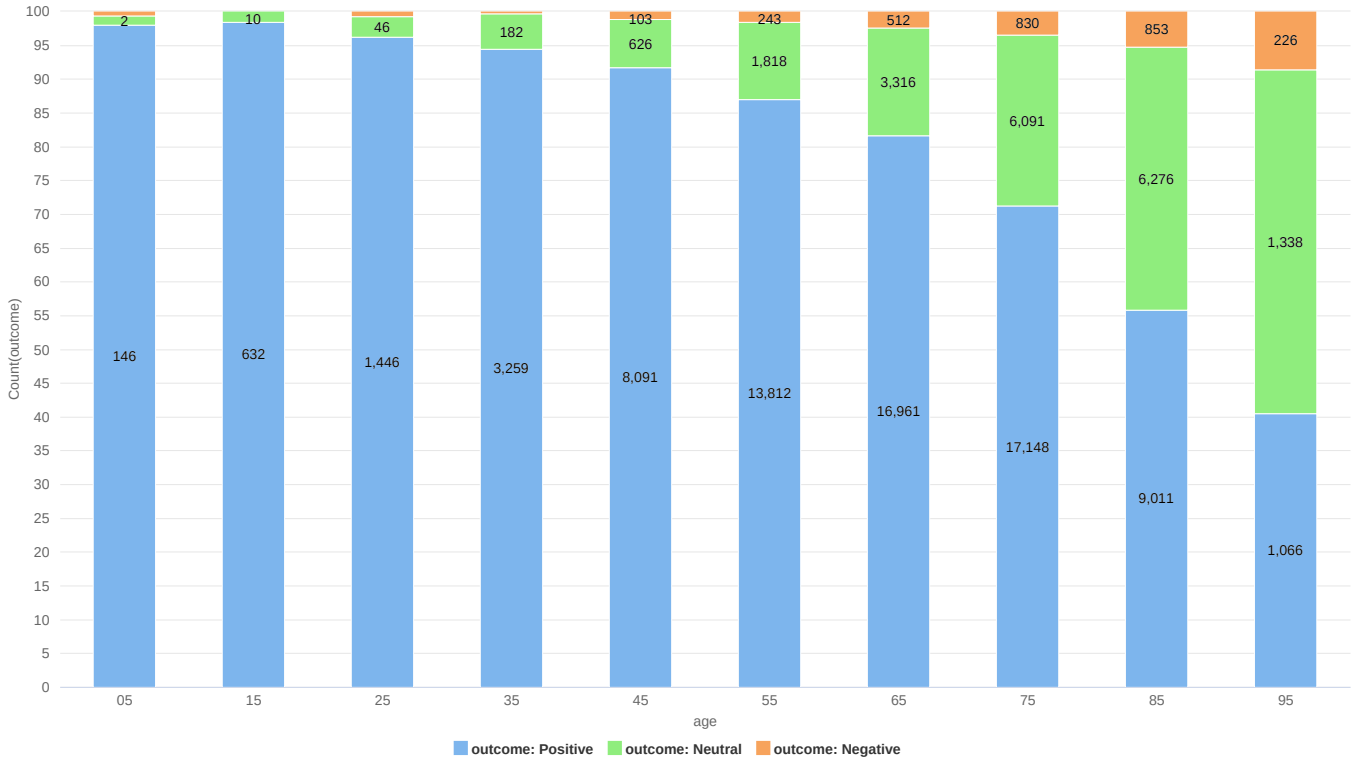
Obrázek 9: Porovnání počtu jednotlivých výsledku hospitalizace v rámci pacientů daného pohlaví



Obrázek 10: Porovnání poměrů jednotlivých výsledku hospitalizace v rámci pacientů daného pohlaví (y-osa značí procenta z celkového počtu daných pacientů)



Obrázek 11: Porovnání počtu jednotlivých výsledku hospitalizace v rámci pacientů daného věku (hodnoty na x-ové osi reprezentují středy intervalů, např. 15 reprezentuje všechny pacienty ve věku 10-20 atd.)



Obrázek 12: Porovnání poměrů jednotlivých výsledku hospitalizace v rámci pacientů daného věku (y-osa značí procenta z celkového počtu daných pacientů, hodnoty na x-ové osi reprezentují středy intervalů, např. 15 reprezentuje všechny pacienty ve věku 10-20 atd.)

Prediktivní model

Při klasifikaci jsme se zaměřili na 3 třídy: `Positive`, `Neutral`, `Negative`. Negativních výsledků bylo nejmíň (viz poslední řádek Tabulky 10), proto výsledný dataset tvoří jen 8370 záznamů. Dataset byl rozdělen následovně: 80 % pro trénování, 20 % pro testování. Byl zvolen prediktivní model Deep Learning, s výchozím nastavením parametrů. Přesnost predikce byla 46.59 % (viz Tabulka 11). Významnosti atributů pro predikci jsou na Tabulce 12. Z významnostních váh hodnot atributů v daném modeli je vidět, že věk měl největší dopad na výsledek hopsitalizace.

Tabulka 11: Přesnost predikce výsledku hospitalizace na základě demografických údajů

accuracy: 46.59%

	true Unknown	true Positive	true Neutral	true Negative	true Still Under Care	class precision
pred. Unknown	0	0	0	0	0	0.00%
pred. Positive	0	326	139	155	0	52.58%
pred. Neutral	0	105	214	163	0	44.40%
pred. Negative	0	127	205	240	0	41.96%
pred. Still Under Care	0	0	0	0	0	0.00%
class recall	0.00%	58.42%	38.35%	43.01%	0.00%	

Tabulka 12: Významnosti hodnot atributů pro predikci

attribute	weight ↓
age.95	1
age.15	0.965
age.35	0.891
age.85	0.864
age.25	0.857
age.45	0.827
age.75	0.792
age.55	0.777
race.Hispanic	0.762
race.AfricanAmerican	0.753
age.05	0.726
race.Asian	0.721
race.Other	0.717
race.Caucasian	0.706
age.65	0.689
gender.Female	0.671
gender.Male	0.605

Korelační analýza

Výstup operátoru *Set Role* (tedy nastavení atributu pro klasifikovaci) byl převeden na operátor *Weight by Chi Squared Statistic*, který spočte testovací kritérium pro test dobré shody (korelační analýza kategorických atributů) v rámci dvojic věk-výsledek, rasa-výsledek, pohlaví-výsledek. Hodnoty tohoto kritéria byly vyčteny z tabulky výsledků této analýzy a jsou nasledovné (čím vyšší hodnota, tím vyšši korelace):

- pohlaví – 427.215
- rasa – 480.919
- věk – 9 563.264

Tyto výsledky jsou konzistentní s váhami prediktivního modelu i s explorativní analýzou. Není překvapením, že pacienti s vyšším věkem mějí s cukrovkou větší potíže. Zajímavým je ale fakt, že ženy v průměru mají horší výsledky hospitalizace. O tom, jestli diabetes vplýva jinak na pacienty různých ras nemůžeme udělat dostatečně přesné závěry, protože dataset neobsahuje porovnatelně velké vzorky pro zástupce různých etnických skupin.