UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Využití bayesovské sítě pro predikci prognózy leukemických pacientů

Katedra matematické analýzy a aplikací matematiky
Vedoucí diplomové práce: **Tomáš Fürst**
Vypracoval(a): **Bc. Štěpánka Matuštíková**
Studijní program: N1103 Aplikovaná matematika
Studijní obor Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2021

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Štěpánka Matuštíková

Název práce: Využití bayesovské sítě pro predikci prognózy leukemických pacientů

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Tomáš Fürst

Rok obhajoby práce: 2021

Abstrakt: Bayesovské sítě jako prostředek predikce prognózy v oblasti medicíny mohou představovat vhodnější alternativu ke klasickým metodám jako je logistická regrese či prognostické indexy. Práce popisuje bayesovské sítě, jejich konstrukci a využití, a to zejména ke klasifikaci událostí a predikci lékařské prognózy. Popsaná teorie je využita při vytvoření bayesovské sítě pro predikci prognózy pacientů s folikulárním lymfomem v České republice na základě vybraných dat od Kooperativní lymfomové skupiny ČR. Výsledky vytvořené sítě a její vlastnosti jsou porovnány s klasickými metodami, které se v této oblasti využívají.

Klíčová slova: Bayesovské sítě, pravděpodobnostní grafické modely, logistická regrese, folikulární lymfom, PRIMA-PI

Počet stran: 100

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Štěpánka Matuštíková

Title: Assessing the prognosis of leukemia patients by means of Bayesian networks

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Tomáš Fürst

The year of presentation: 2021

Abstract: Bayesian networks as a tool for assessing the prognosis of patients may represent a more suitable alternative to classical prognostic methods such as logistic regression or various prognostic indices. The Thesis describes Bayesian networks, their construction, and their use, especially as a predictor of early disease progression. The network is used for the prediction of early progression of patients suffering from follicular lymphoma. The data come from the registry of the Czech Cooperative Lymphoma Group. The performance of the Bayesian network is compared with the standard logistic regression and with the PRIMA prognostic index traditionally used in this area.

Key words: Bayesian networks, probabilistic graphical models, logistic regression, follicular lymphoma, PRIMA-PI

Number of pages: 100

Number of appendices: 1

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana Tomáše Fürsta, a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Poděkování

Ráda bych poděkovala svému vedoucímu diplomové práce Tomáši Fürstovi za možnost pracovat na tak zajímavém projektu a taky za jeho přínosné připomínky. Také bych ráda poděkovala prof. Vítovi Procházkovi za poskytnutá data a možnost se tak zapojit do účasti tohoto projektu na kongresu Americké Hematologické společnosti.

Seznam tabulek

Seznam obrázků

Úvod

Tato práce popisuje využití matematických metod ke stanovení prognózy pacientů v České republice trpící folikulárním lymfomem. Ve spolupráci s prof. MUDr. Vítem Procházkou, Ph.D, a Kooperativní lymfomovou skupinou ČR máme možnost analyzovat data pacientů a aplikovat teorii bayesovských sítí na vytvoření predikčního modelu pro pacienty trpící touto nemocí.

„Folikulární lymfom typicky postihuje lymfatické uzliny. Toto nádorové onemocnění představuje v České republice druhý nejčastější maligní lymfom a pozornost si zaslouží také pestrostí klinických projevů a velmi rozdílnou prognózou nemocných."[1] Zhruba 20 % lidí s touto diagnózou dostane progresi, relaps nemoci, nebo zemře do 2 let od diagnózy. Proto je potřebné tyto pacienty identifikovat při počáteční léčbě ke stanovení správné terapeutické strategie.

Lékaři v této oblasti využívají klasické metody jako je logistická regrese a také speciálně navržené prognostické indexy, které jsou oblíbeny díky své jednoduchosti. Navrhujeme použití alternativní metody bayesovských sítí. Bayesovské sítě jsou grafickými reprezentacemi znalostí s intuitivní strukturou a parametry. Patří do skupiny pravděpodobnostních grafických modelů a jsou hojně využívány v mnoha odvětví a pro různé využití[5], jako například umělá inteligence[2], rozpoznávání vzorců[3], klasifikace[2] a zpracování obrazu[4].

Hlavní vlastností bayesovských sítí, díky kterým zaznamenaly zvýšenou popularitu v posledních letech jsou, že je jejich matematický základ důsledně ověřen, přirozeně se dokáží vypořádat s nejistotou, kterou modelují pomocí sdruženého pravděpodobnostního rozdělení. A díky jejich grafickému zobrazení jsou pochopitelné a využívají lokálnost v reprezentaci znalostí a během inference. Další

výhodou je jejich využití pro predikci nebo také pro deskriptivní modely.[6]

Práce je rozdělena na teoretickou část a praktickou část. V teoretické části je cílem:

- Shrnout klasické prognostické metody - Konkrétně model logistické regrese, který je stručně představen v kapitole 1.2.1. Dále také model PRIMA prognostický index, což je jednoduchý rozhodovací strom, který se využívá k identifikaci úrovně rizika pacienta. Tento index je představen v kapitole 1.2.2.

- Představit bayesovské sítě:

  - vytváření sítí - Tato část je rozdělena na 2 části, a to na vytváření struktury sítě (kapitola 1.3.3), a stanovení parametrů sítě k vybrané struktuře (kapitola 1.3.4).
  - inferenci - Využití bayesovské sítě k inferenci je předvedeno na ilustračním příkladě v kapitole 1.3.2.
  - predikci - Využití a porovnání bayesovské sítě k predikci vybraných dat je ukázáno v praktické části této práce.

Praktická část je věnována zpracováním vybraných dat, které nám byly poskytnuty p. Procházkou. Tato data jsou použita ke splnění cílů praktické části, kterými jsou:

- Sestavit model:

  - Logistické regrese - Tento model je představen v kapitole 2.2.
  - PRIMA-PI - Tento model je ukázán v kapitole 2.3.
  - Bayesovské sítě - Finální model, ale i jiné typy sítí jsou představeny v kapitole 2.4.

- Srovnat modely při predikci EFS24 (Event-free survival at 24 months from diagnosis) na vybraných datech - Toto srovnání lze najít v kapitole 2.16.

Motivace

Tato práce je zaměřená na predikci prognózy pacientů s folikulárním lymfomem v České republice. „Lymfom je obecné označení pro nádorové onemocnění lymfatického systému. Jako lymfomy se označují nádory, které pocházejí z jednoho druhu bílých krvinek, nazývaného lymfocyty". [7] Více o folikulárním lymfomu se lze dočíst v [1]. Tato data byla poskytnuta Českou Kooperaticní lymfomovou skupinou („Jedná se o sdružení lékařů a dalších pracovníků zabývajících se diagnostikou, léčbou a výzkumem v oblasti maligních lymfomů."[8]) Cílem je analyzovat tato data a nalézt vhodné modely k predikci prognózy proměnné EFS24 (Event-free survival 24), která říká, zda daný pacient do dvou let (24 měsíců): zemřel, nebo prodělal progresi (šíření, zhoršení či růst nádoru), nebo relaps (opětovný výskyt symptomů nemoci po období zlepšení). Analyzovaný datový soubor obsahuje dalších 17 proměnných, které jsou v následující tabulce 1 popsány k seznámení čtenáře s touto problematikou.

Proměnná	Vysvětlení
Pohlaví	pohlaví - žena, muž
Věk	věk v letech při 1. diagnóze
Celkové příznaky	pozorovatelné systémové příznaky pacienta (úbytek hmotnosti, horečky apod.): Ano, Ne
Velikost tumoru	velikost největšího tumoru: 0-10+
Stupeň lymfomu	stupeň lymfomu - FL1, FL2, FL1-2, FL3A
Kostní dřeň	infiltrace kostní dřeně lymfomem: Ano, Ne
Postižené extranodální lokalizace	počet mimo uzlinových výskyty lymfomu: 0-2+
Nodální lokalizace	počet uzlinových výskytů lymfomu: 0-9+
Klinické stádium	stádium nemoci podle Ann Arbor klasifikace: I, II, III, IV
Performance status dle ECOG	status dle ECOG klasifikace: 0 až 4
Chemoterapie	typ provedené chemoterapie: R-chop, R-cop, Bendamustine, Intensive, Fludarabine
B2m	hodnota Beta-2-mikroglobulin z krevního vzorku
Leukocyty	hodnota leukocytů z krevního vzorku
Lymfocyty	hodnota lymfocytů z krevního vzorku
Trombocyty	hodnota trombocytů z krevního vzorku
Hemoglobin	hodnota hemoglobinu z krevního vzorku
LDH vyšší než norma	hodnota LDH (laktátdehydrogenáza) z krevního vzorku vyšší než norma: Ano, Ne
EFS24	událost (smrt, progrese, relapse) do dvou let od diagnózy: Ano, Ne

Tabulka 1: Název a vysvětlení proměnných vyskytujících se v této práci

Kapitola 1

Teoretická část

V této kapitole se stručně seznámíme s klasickými metodami využívanými k predikci prognózy pro leukemické pacienty. Přesněji s modelem logistické regrese a PRIMA prognostickým indexem. Velká část této kapitoly se bude zabývat hlavním tématem této práce, bayesovským sítím, u kterých se seznámíme hlavně s jejich konstrukcí z dat a využitím pro klasifikaci. Tyto znalosti nám poslouží k vytvoření a porovnání modelů ke stanovení prognózy pacientů s folikulárním lymfomem na poskytnutých datech.

1.1. Klasifikační úloha a její hodnocení

Jako první je potřeba se seznámit s typem úlohy, kterým se tato práce zabývá, a to je klasifikační úloha. Tato úloha klasifikuje vybranou proměnnou do určité z hodnot (kategorií), kterých může nabývat, na základě předložených hodnot vysvětlujících proměnných, tedy těch ostatních. Základním typem je binární klasifikace, kde výsledek je 0 nebo 1, nebo také ANO nebo NE. Takovou to úlohu si můžeme představit jako otázku, jak podle výšky, váhy a délky vlasů dokážeme posoudit zda popisovaná osoba je muž či žena. S tímto typem úlohy se budeme zabývat. Čerpáno z [9].

Zajímá nás tedy klasifikace dichotomické proměnné. Ostatní proměnné mohou být kategorické či kvantitativní. Některé algoritmy stanoví přímo třídu klasifikace, např. algoritmus k-nejbližších sousedů, a některé modelují hodnotu mezi

0 a 1, jako logistická regrese. Při této možnosti se na základě vybraného algoritmu a hodnot vysvětlujících proměnných vypočítá hodnota mezi 0 a 1, která představuje odhad pravděpodobnosti, že pozorování nabývá hodnoty 1. Poté se stanoví práh, podle kterého se vypočtené hodnoty klasifikují jako 0 nebo 1. Díky tomu se může sestavit matice záměn, která je zobrazena na obrázku 1.1, kde vidíme kontingenční tabulku pro skutečné hodnoty a pro predikované hodnoty. V tabulce jsou označeny množství pozorování, které odpovídají výsledkům modelu a skutečnosti. S těmito množstvími se pracuje v následujících měřítkách.

		Skutečné hodnoty	
		1	**0**
Predikované hodnoty	**1**	True Positives	False Positives
	0	False Negatives	True Negatives
	∑	Positives	Negatives

Tabulka 1.1: Matice záměn

V praktické části jsou použita tato porovnávající měřítka:

- Přesnost, accuracy, je celkové množství správně identifikovaných na celkovém množství pozorování:

$$Accuracy = \frac{TP + TN}{P + N} \tag{1.1}$$

- Citlivost, senzitivita, také true positive rate (TPR), která vyjadřuje podíl správně pozitivně identifikovaných na celkovém počtu všech skutečně pozitivních:

$$Sesitivity = TPR = recall = \frac{TP}{TP + FN} \tag{1.2}$$

- Specifičnost, specificity, také true negative rate (TNR), která vyjadřuje poměr správně negativně identifikovaných na celkovém počtu všech skutečně

negativních:

$$Specificity = TNR = \frac{TN}{FP + TN} \tag{1.3}$$

- Precision, která vyjadřuje podíl správně pozitivně identifikovaných na celkovém množství pozitivně označených vybraným modelem:

$$Precision = \frac{TP}{TP + FP} \tag{1.4}$$

- Míra falešně pozitivních výsledků, FPR, která vyjadřuje podíl falešně pozitivních na celkovém počtu všech skutečně negativních:

$$FPR = \frac{FP}{FP + TN} = 1 - specificity \tag{1.5}$$

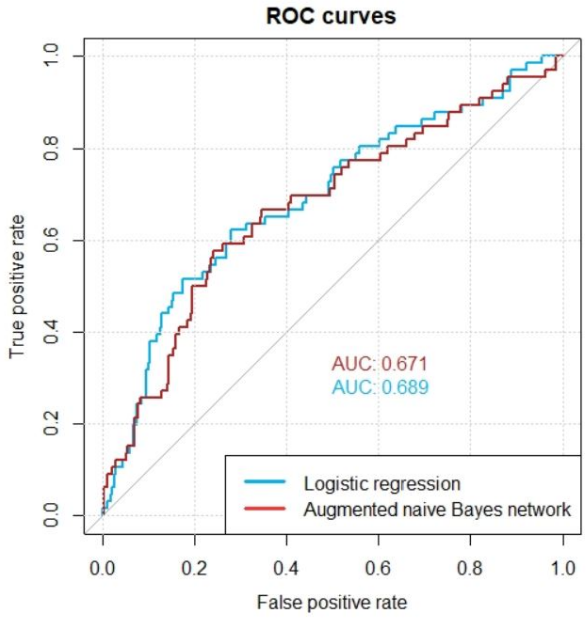
- F1 skóre

$$F1 = 2 * \frac{precision * recall}{precision + recall} = 2 * \frac{precision * sensitivity}{precision + sensitivity} \tag{1.6}$$

Jak jde z jasného popisu vidět, tak každé z těchto měřítek má různé výhody a nevýhody, především nám každé z nich poskytuje lehce odlišnou informaci o modelu. Obecně je cíleno k vysoké přesnosti, ta u některých případů, kdy nejsou klasifikované třídy symetricky rozděleny, může způsobit nesprávné posuzování modelů. V některých úlohách je důležitější klasifikace pozitivních jedinců, a tak se pozornost upírá spíše k citlivosti, která je ale negativně korelována se specifičností v závislosti na volbě prahové hodnoty. Vždy tedy záleží na povaze úlohy, a je dobré se dívat na více měřítek.

Dále je používán ROC graf (receiver operating characteristics), poměrně běžně užívaný v lékařském rozhodování a jiných doménách datové analýzy a strojového učení. ROC graf je dvoudimenzionální graf, kde je hodnota citlivosti vynesena na osu Y a hodnota míry falešně pozitivních výsledků je vynesena na osu X. U algoritmů, u kterých je výstupem 0 nebo 1 je výstupem pouze jedna matice záměn, a po vypočítání hodnoty citlivosti a FPR je výstupem jeden bod v prostoru ROC grafu. V případě algoritmů, u kterých je výstupem hodnota mezi 0

a 1 jsou vypočítány zmiňované hodnoty TPR a FPR pro matice záměn pro prahové hodnoty od 0 do 1. Vzniká křivka z bodu [0;0] do bodu [1;1]. Bod [0;0] na vzniklé křivce představuje klasifikátor, který všechny hodnoty určil jako 0. Bod [1;1] naopak určil všechny jako 1. Bod [0;1] je nejlepší výsledek, jelikož vyhodnotil všechny hodnoty správně. Citlivost se v tomto bodě rovná 1, tedy False Negatives neobsahuje žádné pozorování. FPR se rovná 0, takže False positives neobsahuje žádné pozorování. S tímto grafem je spojena hodnota AUC, area under the ROC curve, která vypočítá plochu pod křivkou. Tato hodnota slouží k porovnání klasifikátorů. Nejlepší možná hodnota je 1, která jde z bodu [0;0] do [0;1] a poté do [1;1]. Hodnota 0,5 udává diagonální křivku z bodu [0;0] do [1;1], která odpovídá náhodnému klasifikátoru. Příklad takovéto ROC křivky s vypočítanou hodnotou AUC je uvedena na obrázku 1.1. Více ohledně ROC grafů lze nalézt v [9].



Obrázek 1.1: ROC křivka

1.2. Klasické modely

Velmi využívanými metodami ke stanovení prognózy leukemických pacientů jsou dnes: model logistické regrese a jednodušší model PRIMA prognostický index (PRIMA-PI). Tyto metody si ve stručnosti představíme.

1.2.1. Logistická regrese

Logistická regrese je celosvětově využívaná metoda ke klasifikaci v mnoha oborech. Je to stálá metoda, které dosahuje dobrých výsledků i v porovnáním s novými efektivními modely. Slouží tak často jako tzv. benchmark, takže se často využívá k porovnání. Čerpáno z [10].

Šance a poměr šancí

Pravděpodobnost událostí se dá vyjádřit mnoha způsoby, šance události vyjadřuje podíl očekávaných četností, že událost nastane, oproti očekávaným četnostem, že událost nenastane. Tento poměr je využíván ke přeměně lineárního regresního modelu do modelu logistické regrese. Pokud p je pravděpodobnost události, tak O je šance této události.

$$O = \frac{p}{1-p} \tag{1.7}$$

Pokud je O větší než 1, tak je pravděpodobnost $p > 0.5$, a naopak. Hlavním využití šance je při porovnávání dvou dichotomických proměnných, což je ilustrováno na následujícím příkladu.

Příklad 1.2.1 *Máme dvě proměnné: zdravotní stav a kuřák. Tyto hodnoty nabývají dvou kategorií, jak je zobrazeno v kontingenční tabulce na obrázku 1.2.*

V tabulce je vypočítaná také šance infarktu pro kuřáky, nekuřáky a všechny pozorování. Poměr šanci pro kuřáky a nekuřáky je $\frac{3.25}{0.48} = 6.77$*. Takže lze říci, že kuřák má 6,77 krát větší šanci, že dostane infarkt než nekuřák.*

	Kuřák	Nekuřák	∑
Infarkt	39	16	55
Zdravý	12	33	45
∑	51	49	**100**
Šance Infarktu	39/12=3.25	16/33=0.48	55/45=1.22

Tabulka 1.2: Kontingenční tabulka pro proměnné zdravotní stav a kuřák

Stejně jako pravděpodobnost, tak šance je ohraničená nulou. Na druhou stranu není ohraničená zprava.

Model logistické regrese

Logistická regrese je využívána k modelování střední hodnoty dichotomické závislé proměnné pomocí nezávislých náhodných proměnných $X_1, \ldots, X_k$. Také je využívána ke klasifikaci. Výstup lineární regrese může nabývat jakékoliv reálné hodnoty. Pro klasifikaci je potřebná hodnota mezi 0 a 1. Proto v modelu logistické regrese je pravděpodobnost, která je ohraničená nulou a jedničkou, transformována do šance, čímž se odstraní horní hranice a transformováním logaritmem se odstraní i dolní hranice. Poté je výsledek modelován lineární funkcí vysvětlujících proměnných. Logistický model vypadá následovně:

$$log\Big[\frac{p_i}{1-p_i}\Big] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \tag{1.8}$$

kde p_i je pravděpodobnost, že $y_i = 1$. Hodnoty vysvětlovaných proměnných, tedy x_i mohou být kvantitativní nebo to mohou být umělé proměnné, které nabývají hodnot 0 nebo 1. Z rovnice se dá vyjádřit p_i:

$$p_i = \frac{1}{1 + exp(-\alpha - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_k x_{ik})}. \tag{1.9}$$

Výsledné číslo bude pro jakékoliv hodnoty β_i a x_i mezi 0 a 1, což je cílená vlastnost. Parametry modelu jsou vypočítány metodou maximální věrohodnosti, což znamená nalézt množinu parametrů, pro které je pravděpodobnost pozorovaných dat největší.

Model logistické regrese je velmi široce popsán v literatuře, proto se jím dál nebudeme zabývat.

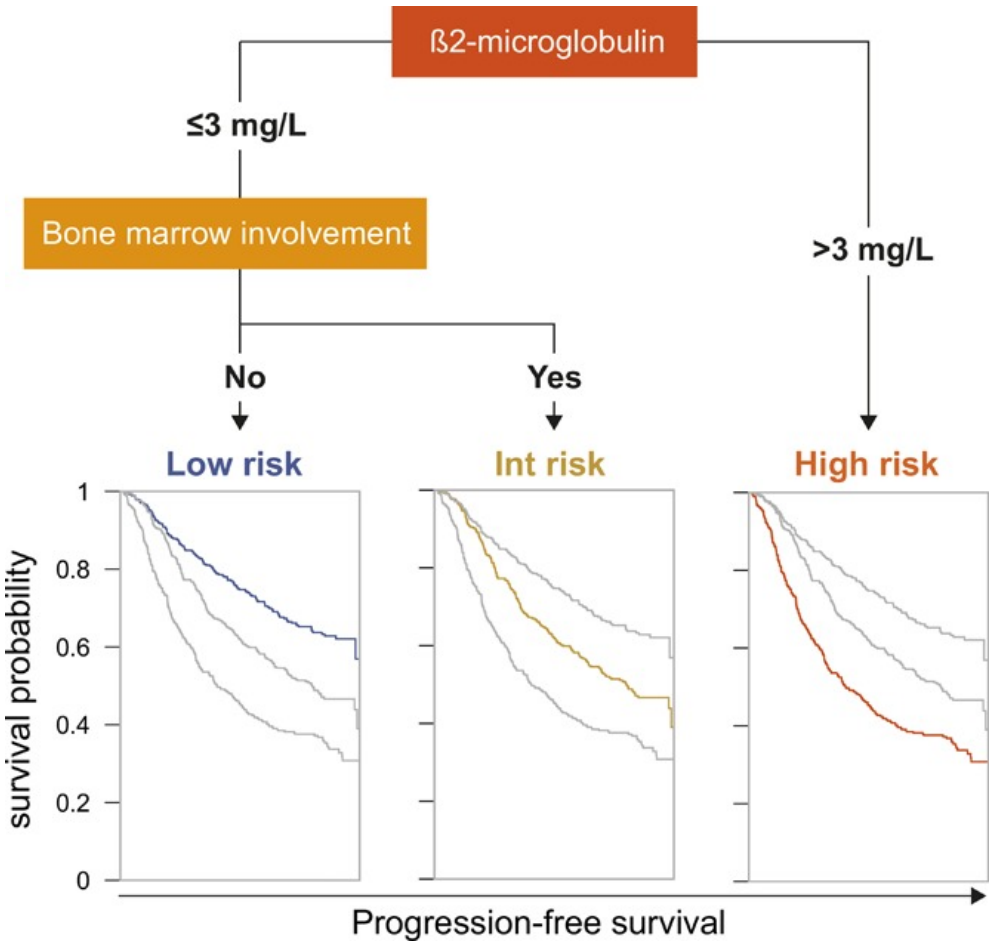
1.2.2. PRIMA-PI

PRIMA - prognostic index, popsán v [11], udává jednoduchým způsobem kategorii rizika pacienta pomocí dvou proměnných, a to B2m a kostní dřeně. Model rozřazuje pacienty do kategorie: nízké, střední a vysoké riziko. U těchto kategorií je pak stanovena pravděpodobnost výskytu události, která je stanovena podle kontrolní skupiny jako její frekvence.

Model je postaven jako rozhodovací strom s kořenem proměnné B2m. Pro výběr těchto dvou proměnných byla provedena analýza dat, která určila 5 statisticky významných proměnných: pohlaví, LDH, kostní dřeň, hemoglobin a B2m. Tyto proměnné byly uvažovány do stavby rozhodovacího stromu o maximální hloubce 2. Jako optimální vyšel právě daný rozhodovací strom s B2m a kostní dření, který je zobrazen na obrázku 1.2.

Výpočet kategorie rizika probíhá jednoduše tak, že se podíváme na pacientovu hodnotu B2m a pokud je vyšší než 3, tak pacient spadá do kategorie vysokého rizika. Pokud je rovna nebo nižší než 3, tak se podíváme na hodnotu kostní dřeně. Pokud má pacient postiženou kostní dřeň tak spadá do středního rizika, a pokud ji nemá postiženou tak spadá do nízkého rizika. Takže pacienti s hodnotou B2m vyšší než 3 jsou ve vysokém riziku, pacienti s hodnotou B2m nižší nebo rovna 3 a postiženou kostní dření jsou ve středním riziku a pacienti s hodnotou B2m nižší nebo rovna 3 a nepostiženou kostní dření jsou v nízkém riziku.

Tento index je podle výzkumů korelován s proměnnou EFS24. Výzkum dospěl k závěru, že PRIMA prognostický index je stejně vypovídající jako do té doby nejvyužívanější index FLIPI a výkonnější než index LDH+B2m. Jako největší výhoda tohoto indexu je jeho jednoduchost v rutinní klinické praxi, která plyne z potřeby znát hodnoty pouze 2 proměnných.



Obrázek 1.2: Rozhodovací strom využívaný ke stanovení PRIMA-PI, [11]

1.3. Bayesovské sítě jako pravděpodobnostní grafický model

Pravděpodobnostní grafické modely umožňují zachycení nejistoty, která se vyskytuje ve většině systémech reálných aplikací, které se snažíme pochopit. Práci s nejistotou se většinou nevyhneme, jelikož většiny stavů systému nelze přesně změřit. Buď je opravdový stav nezjistitelný nebo hodnoty mohou být nepřesné kvůli chybě měření. [12]

Tyto modely umožňují transformovat složitý systém do kompaktnější formy díky pochopení nezávislostí a závislostí vztahů mezi proměnnými v systému. Tato schopnost dále usnadňuje proces inference. Modely jsou aplikované v různých odvětví, jako například: diagnóza, expertní systémy, plánovací systémy, data analýza a kontrola. Modely jsou sestaveny z kvalitativní a kvantitativní komponenty. Kvalitativní komponenta představuje strukturu grafu, která zobrazuje unikátním přístupem podmíněné nezávislosti mezi proměnnými. Kvantitativní komponenta představuje množinu parametrů v modelu příslušící vybrané struktuře grafu.[13] V této kapitole bylo vycházeno zejména z [12], [14], [15] a [16].

1.3.1. Graf

Graf $\mathcal{G} = (\mathcal{X}, \Psi)$ zde představuje reprezentaci struktury dat. Je to soubor uzlů a soubor hran, kde jsou uzly propojeny hranami. Soubor uzlů představuje množinu náhodných veličin vstupující do modelu, $\mathcal{X} = \{X_1, ..., X_k\}$. Náhodné veličiny jsou veličiny, které vlivem náhodných činitelů nabývají různých hodnot. Hrany mohou být orientované nebo neorientované a tvoří soubor dvojic Ψ, kde dvojice mohou být $X_i \leftarrow X_j$, $X_i \rightarrow X_j$ nebo $X_i - X_j$, pro $X_i, X_j \in \mathcal{X}, i < j$. Jelikož strukturou bayesovské sítě je orientovaný acyklický graf, tak dále budeme předpokládat pouze první dvě možnosti. Směrové šipky reprezentují přímé stochastické závislosti mezi proměnnými. Jestliže nejsou mezi dvěma proměnnými šipky, tak to znamená, že jsou proměnné buď marginálně nezávislé nebo podmíněně nezávislé. Pokud $X_i \leftarrow X_j$, tak se X_i označuje jako potomek X_j

a X_j jako rodič X_i. Stupeň uzlu označuje celkové množství hran, na kterých se podílí. Stupeň grafu je nejvyšší hodnota stupně uzlu. [12]

Orientovaný acyklický graf (DAG)

Orientovaný acyklický graf je tvořen pouze orientovanými hranami a žádná cesta netvoří cyklus. Cyklus v grafu $\mathcal{G}$ je přímá cesta $X_1, \ldots, X_k$, kde $X_1 = X_k$. Graf je acyklický pokud neobsahuje žádné cykly. [12]

1.3.2. Představení bayesovských sítí

Bayesovské sítě (BN) jsou reprezentovány pomocí DAG. Jelikož spojité proměnné v našem datovém souboru v praktické části nesplňují předpoklad normality, tak se dále budeme věnovat pouze diskrétním bayesovským sítím. Jak napovídá název, jsou sítě založeny na bayesovské statistice, která vychází z Bayesova pravidla. Čerpáno z [12] a [14].

Definice 1.3.1 (Podmíněná pravděpodobnost [14]) *Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor a $A, B \in \mathcal{A}$, $P(B) > 0$. Podmíněnou pravděpodobnost jevu A za podmínky B definujeme vztahem*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Věta 1.3.1 (o násobení pravděpodobností [14]) *Pro libovolný $n + 1$ jevů $A_0, \ldots, A_n$ takových, že $P(A_0 A_1 \ldots A_{n-1}) > 0$, platí*

$$P(A_0 A_1 \ldots A_n) = P(A_0) P(A_1|A_0) \ldots P(A_n|A_0 A_1 \ldots A_{n-1}).$$

Věta 1.3.2 (o celkové pravděpodobnosti [14]) *Je-li $P(\cup_n B_n = 1)$, kde $\{B_n\}$ je konečná nebo spočetná posloupnost navzájem se vylučujících jevů, je-li $P(B_n) > 0$ pro všechna n, a je-li $A \in \mathcal{A}$, potom*

$$P(A) = \sum_n P(A|B_n) P(B_n).$$

Věta 1.3.3 (Bayesova [14]) *Za předpokladu předchozí věty a za předpokladu* $P(A) > 0$ *platí*

$$P(B_m|A) = \frac{P(A|B_m)P(B_m)}{\sum_n P(A|B_n)P(B_n)} \qquad \textit{pro všechna } m.$$

V bayesovské statistice označujeme jmenovatele z Bayesova pravidla jako normalizační faktor a $P(B_m)$ jako apriorní pravděpodobnost, která nám dává možnost přidat apriorní znalost do modelu. Využití bayesovských sítí je zobrazeno na následujícím ilustračním příkladu (vlastní tvorba). Za ním následuje využitá teorie.

Příklad 1.3.1 *Budeme sledovat pravděpodobnost uklouznutí a způsobení zlomeniny uklouznutím na základě něčí obratnosti a také toho, zda pršelo nebo ne. Máme 4 náhodné proměnné: Déšť, Obratnost, Uklouznutí a Zlomenina, každá z těchto proměnných nabývá 2 nebo 3 hodnot: déšť - pršelo nebo nepršelo; obratnost - špatná, průměrná, dobrá; uklouznutí - ano, ne; zlomenina - ano, ne. Jejich pravděpodobnostní rozdělení je zobrazeno pomocí diskrétní bayesovské sítě na obrázku 1.3. Každá proměnná v bayesovské síti je definována pomocí CPT, tedy tabulek podmíněného pravděpodobnostního rozdělení (conditional probability tables), které jsou zobrazeny na obrázku. Na tomto příkladě si předvedeme využití bayesovské sítě při výpočtu sdruženého pravděpodobnostního rozdělení a při inferenci.*

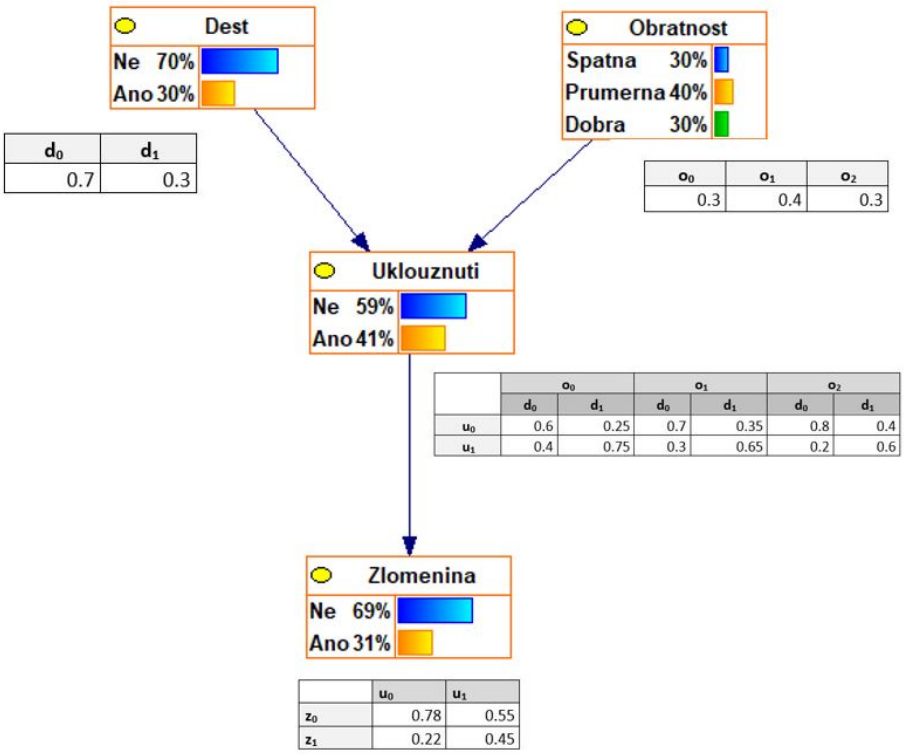
Sdruženého pravděpodobnostního rozdělení:

- *Pomocí věty 1.3.1:*

$$P(D, O, U, Z) = P(D)P(O|D)P(U|D, O)P(Z|D, O, U).$$

- *S využitím řetěz. pravidla pro BN (viz. definice 1.3.4) a využitím struktury zadané bayesovské sítě:*

$$P(D, O, U, Z) = P(D)P(O)P(U|D, O)P(Z|U).$$



d_0	d_1
0.7	0.3

o_0	o_1	o_2
0.3	0.4	0.3

	o_0		o_1		o_2	
	d_0	d_1	d_0	d_1	d_0	d_1
u_0	0.6	0.25	0.7	0.35	0.8	0.4
u_1	0.4	0.75	0.3	0.65	0.2	0.6

	u_0	u_1
z_0	0.78	0.55
z_1	0.22	0.45

Obrázek 1.3: Ilustrační příklad - Bayesovská síť pro zlomeninu

U první možnosti vidíme, že by bylo potřeba vypočítat jednotlivé komponenty výpočtu, které by byly složitější než druhá možnost, protože u této možnosti můžeme všechny údaje dostat z dat z CPT dané bayesovské sítě. Například pro pravděpodobnost, že bude pršet, mám špatnou obratnost, uklouznu a zlomím si nohu, je výpočet následující:

$$P(d_1, o_0, u_1, z_1) = P(u_1|o_0, d_1)P(d_1)P(o_0)P(z_1|u_1) = 0.75*0.3*0.3*0.45 = 0.0304$$

Inference

Vypočítání pravděpodobnosti vybrané proměnné na základě předložených důkazů: jaká je pravděpodobnost, že uklouznu, pokud bude prše, mám špatnou obratnost a zlomím si nohu.

$$P(u_1|d_1, o_0, z_1) = \frac{P(u_1, z_1, d_1, o_0)}{P(d_1, o_0, z_1)}$$

Čitatel je vypočítán výše a jmenovatel se vypočte pomocí věty 1.3.2 a pomocí

struktury sítě následovně:

$$P(z_1, d_1, o_0) = \sum_{i=\{0,1\}} P(z_1|u_i)P(u_i|d_1, o_0)P(d_1)P(o_0) =$$

$$= 0.22 * 0.25 * 0.3 * 0.3 + 0.45 * 0.75 * 0.3 * 0.3 = 0.0353$$

Celkový výpočet je tedy:

$$P(u_1|d_1, o_0, z_1) = \frac{0.0304}{0.0353} = 0.86$$

Pravděpodobnost, že uklouznu, pokud bude pršet, jsem špatně obratná a zlomím si nohu je 86 %. Takže jde vidět, že vliv informace od potomka putuje i k rodiči. Informace o tom, že si zlomím nohu potom ovlivní pravděpodobnost uklouznutí. Pokud bychom zjišťovali pravděpodobnost, že uklouznu na základě toho, že bude pršet a já mám špatnou obratnost, tedy $P(u_1|d_1, o_0)$, tak by pravděpodobnost byla v CPT u uklouznutí a rovnala by se 75 %.

Nyní přidáme hranu z proměnné obratnost do proměnné zlomenina, viz obrázek 1.4, a můžeme sledovat, co se stane.

Sdruženého pravděpodobnostního rozdělení: *Možnost, že bude pršet, mám špatnou obratnost, uklouznu a zlomím si nohu:*

$$P(d_1, o_0, u_1, z_1) = P(z_1|u_1, o_0)P(u_1|o_0, d_1)P(d_1)P(o_0) = 0.63{*}0.75{*}0.3{*}0.3 = 0.0425$$

Inference

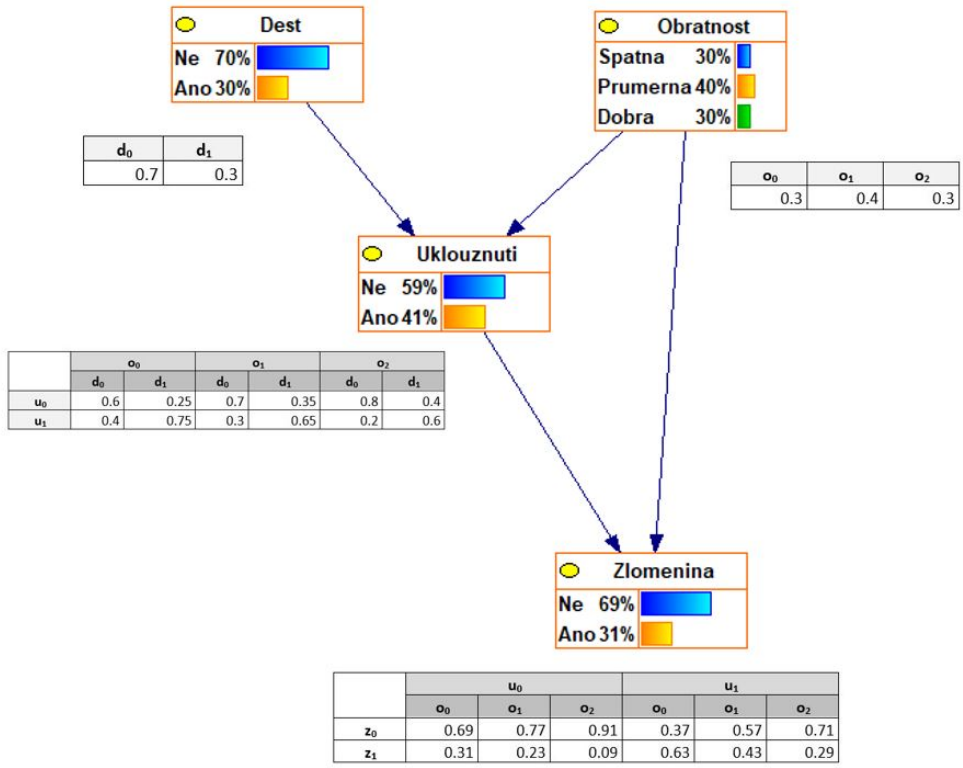
Jaká je pravděpodobnost, že uklouznu, pokud bude prše, mám špatnou obratnost a zlomím si nohu.

$$P(u_1|d_1, o_0, z_1) = \frac{P(u_1, z_1, d_1, o_0)}{P(d_1, o_0, z_1)}$$

Čitatel je vypočítán výše a jmenovatel se vypočte:

$$P(z_1, d_1, o_0) = \sum_{i=\{0,1\}} P(z_1|u_i, o_0)P(u_i|d_1, o_0)P(d_1)P(o_0) =$$

$$= 0.31 * 0.25 * 0.3 * 0.3 + 0.63 * 0.75 * 0.3 * 0.3 = 0.0495$$



d_0	d_1
0.7	0.3

o_0	o_1	o_2
0.3	0.4	0.3

	o_0		o_1		o_2	
	d_0	d_1	d_0	d_1	d_0	d_1
u_0	0.6	0.25	0.7	0.35	0.8	0.4
u_1	0.4	0.75	0.3	0.65	0.2	0.6

	u_0			u_1		
	o_0	o_1	o_2	o_0	o_1	o_2
z_0	0.69	0.77	0.91	0.37	0.57	0.71
z_1	0.31	0.23	0.09	0.63	0.43	0.29

Obrázek 1.4: Ilustrační příklad - Bayesovská síť pro zlomeninu 2

Dohromady: $P(u_1|d_1, o_0, z_1) = \frac{0.0425}{0.0495} = 0.859$

Pravděpodobnost, že uklouznu, pokud bude pršet, jsem špatně obratná a zlomila jsem si nohu je 85.9 %, což je jen o desetinu procenta méně než v předešlém příkladě. Výsledky jsou podobné protože celkové pravděpodobnosti jednotlivých proměnných zůstaly podobné. Rozdíl je ve výpočtu, který se lehce pozměnil na základě změny struktury sítě.

Cílem bayesovské sítě je kompaktnější pochopení sdruženého rozdělení zadaných náhodných proměnných $\mathcal{X} = \{X_1, \ldots, X_n\}$. Sdružené rozdělení dichotomických proměnných by potřebovalo specifikaci $2^n - 1$ čísel, což je velmi náročné na manipulaci a také nemožné ke specifikaci expertem. Statisticky by byla potřeba velké množství dat ke stanovení těchto parametrů robustně. Tyto problémy jsou hlavními motivacemi bayesovských sítí, které využívají nezávislostí mezi proměnnými ke zjednodušení zmiňovaného sdruženého pravděpodobnostního rozdělení.

Definice 1.3.2 (Nezávislost jevů [14]) *Jevy A,B se nazývají nezávislé, jestliže* $P(A, B) = P(A)P(B)$. *V opačném případě mluvíme o jevech závislých.*

Věta 1.3.4 *Jsou-li jevy A,B nezávislé, a je-li navíc* $P(B) > 0$, *pak* $P(A|B) = P(A)$. *[14]*

Definice 1.3.3 (Bayesovská síť) *Bayesovská síť je dvojice* $\mathcal{B} = (\mathcal{G}, P)$, *kde* P *se faktorizuje přes* $\mathcal{G}$, *a kde je* P *specifikováno jako množina podmíněných pravděpodobnostních rozděleních příslušícím k uzlům grafu* $\mathcal{G}$. *Rozdělení* P *se často označuje* $P_{\mathcal{B}}$.

Bayesovské sítě umožňují zjednodušené výpočty sdružených rozděleních a inference pomocí tzv. řetězového pravidla pro bayesovské sítě.

Definice 1.3.4 (Řetězové pravidlo pro bayesovské sítě) *Nechť* $\mathcal{G}$ *je struktura bayesovské sítě pro náhodné proměnné* $X_1, \ldots, X_n$. *Řekneme, že pravděpodobnostní rozdělení* P *se faktorizuje podle* $\mathcal{G}$, *pokud* P *lze vyjádřit jako součin:*

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa_{X_i}^{\mathcal{G}}),$$

kde $Pa_{Xi}^{\mathcal{G}}$ *je množina rodičů uzlu* X_i. *Tato rovnost se nazývá řetězové pravidlo pro BN. Individuální faktory* $P(X_i|Pa_{X_i}^{\mathcal{G}})$ *jsou nazývány podmíněné pravděpodobnostní rozdělení (CPD) nebo lokální pravděpodobnostní modely.*

CPD je v diskrétním případě vyjádřeno pomocí tabulek (CPT), které jsou dány znalostí nebo jsou vypočteny z dat. Jejich velikost záleží na počtu kategorií proměnné a počtu rodičů daného uzlu, číslo parametrů potřebných v těchto tabulkách roste exponenciálně s počtem rodičů. Například, pokud má uzel binární proměnné 4 binární rodiče, tak je potřeba specifikovat $2^4 = 16$ parametrů, pokud má 7, tak je potřeba $2^7 = 128$ parametrů.

Související pojmem s řetězovým pravidlem pro BN je také tzv. Markov blanket, $\mathcal{B}(X_i)$, díky které je definováno, že uzel X_i je nezávislý na všech ostatních uzlech při znalosti jeho $\mathcal{B}(X_i)$, což ostatní uzly dělá nepotřebnými při inferenci. Markov blanket proměnné X_i je definována jako rodiče, potomci a manželé (ostatní rodiče potomků) uzlu X_i.

V bayesovských sítí je možné vypočítat podmíněné i sdružené rozdělení velmi jednoduše a rychle. Grafické znázornění pomáhá při představě o tom, jak daný systém funguje.

Jak bylo řečeno, tak BN jsou tvořeny strukturou grafu a parametry. Obě komponenty je možné stanovit s expertní znalostí a nebo také odvodit z dat $\mathcal{D}$, u kterých se předpokládá, že pozorování jsou IID z P^* (skutečné pravděpodobnostní rozdělení). Cílem je nalézt model $\mathcal{M}^*$, který přesně vystihuje skutečné pravděpodobnostní rozdělení P^*, ze kterého byla data vybrána. Bohužel, tento cíl je obecně nedosažitelný, kvůli výpočetní náročnosti a hlavně kvůli datové množině, která většinou dává pouze hrubou aproximaci skutečného pravděpodobnostního rozdělení. [12] Proto se v praxi konstruuje model $\tilde{\mathcal{M}}$, který nejlépe aproximuje skutečný model $\mathcal{M}^*$. Vyhodnocení nejlepšího modelu záleží na daném cíli, podle kterého se vybírá hodnotící kritérium.

Cíle pro vytváření bayesovských sítí

- Odhad pravděpodobnostního rozdělení - Tento cíl je běžně využíván k pravděpodobnostnímu odvozování (inferenci). Konstruuje se model $\tilde{\mathcal{M}}$ tak, aby $\tilde{P}$ bylo co nejpodobnější P^*.

- Speciální predikční zadání - Cílem je, aby model nejlépe odpovídal na určitý druh pravděpodobnostních dotazů, například predikci hodnoty určité proměnné.

- Objevování znalostí - Cílem je zde nalézt nové znalosti o P^* jako jsou přímé nebo nepřímé závislosti. Rozdíl od odhadu pravděpodobnostního rozdělení je v tom, že cílem je nalézt správný model $\mathcal{M}^*$, ne jen jeho odhad $\tilde{\mathcal{M}}$. Při odhadu pravděpodobnostního rozdělení je pozornost zaměřena na rozdíl P^* a $\tilde{P}$. Bohužel, tento cíl je většinou nedosažitelný kvůli neidentifikovatelnosti (několik struktur může být ve srovnání ekvivalentních). [12]

1.3.3. Struktura bayesovské sítě

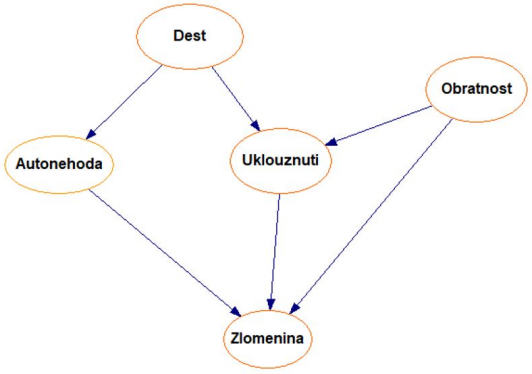
Struktura bayesovské sítě je označována jako kvalitativní komponenta a je velmi důležitou součástí. Nalezení struktury grafu je první ze dvou částí vytvoření bayesovské sítě.

Definice 1.3.5 (Struktura bayesovské sítě) *Struktura bayesovské sítě $\mathcal{G}$ je orientovaný acyklický graf, jehož uzly reprezentují náhodné proměnné $X_1, \ldots, X_n$. Nechť $Pa^{\mathcal{G}}_{X_i}$ jsou rodiče uzlu X_i v $\mathcal{G}$, a $NePotomci_{X_i}$ značí proměnné, které nejsou potomky X_i. Potom pro $\mathcal{G}$ platí následující sada předpokladů podmíněné nezávislosti, nazývané lokální nezávislosti $\mathcal{I}_{\mathcal{L}}(\mathcal{G})$:*

Pro každou proměnnou X_i platí: $(X_i \perp NePotomci_{X_i} | Pa^{\mathcal{G}}_{X_i})$.

Podle obrázku 1.5, kde je vyobrazena rozšířená síť z příkladu1.3.1, a definice 1.3.5 je zřejmý následující vztah:

$$(Autonehoda \perp Uklouznuti | Dest).$$



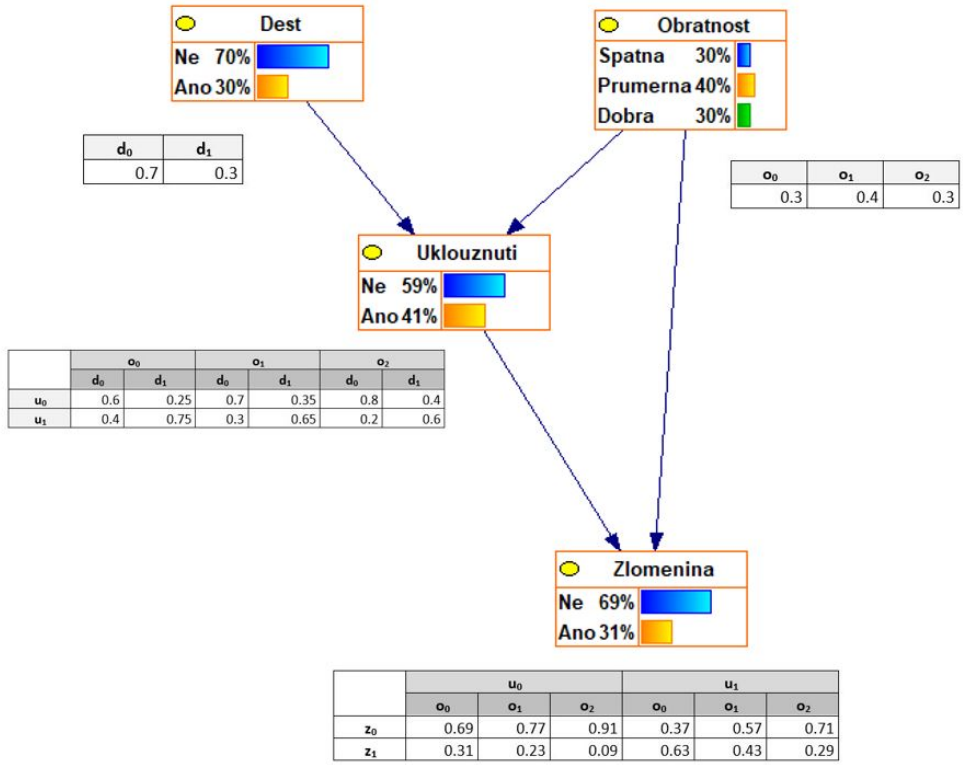
Obrázek 1.5: Jednoduchá bayesovská síť pro představení lokálních nezávislostí

Nalezení struktury bayesovské sítě z dat

Nalezení struktura sítě představuje pro určité úlohy hlavní cíl. Nicméně, tento problém není jednoduchý. Pro účely této práce jsou předpokládaná úplná data, tedy bez chybějících pozorování. Častá je i práce s neúplnými daty v této problematice. Existují zde dvě základní metody učení struktury, které jsou dále popsány,

a to učení struktury na základě omezení vztahů a na základě skórovacích funkcí. Další metody jsou bayesovské průměrování modelu, které lze najít v [12], nebo hybridní modely, které kombinují dva dále rozebírané přístupy. V následující podkapitole bylo vycházeno převážně z [6] a [12].

Pro ilustraci využití metod budeme využívat údaje z příkladu 1.3.1. Vygenerujeme si data *Zlomenina* o 10 000 pozorování podle rozdělení dané CPT z obrázku 1.6. Tato data budeme využívat k nalezení struktury pomocí různých metod.



d_0	d_1
0.7	0.3

o_0	o_1	o_2
0.3	0.4	0.3

	o_0		o_1		o_2	
	d_0	d_1	d_0	d_1	d_0	d_1
u_0	0.6	0.25	0.7	0.35	0.8	0.4
u_1	0.4	0.75	0.3	0.65	0.2	0.6

	u_0			u_1		
	o_0	o_1	o_2	o_0	o_1	o_2
z_0	0.69	0.77	0.91	0.37	0.57	0.71
z_1	0.31	0.23	0.09	0.63	0.43	0.29

Obrázek 1.6: Bayesovská síť a CPD tabulky podle kterých provedeme vygenerování dat

Určení struktury sítě na základě omezení vztahů

V tomto přístupu je cílem zachytit ve struktuře sítě všechny nezávislosti v doméně. Probíhají zde tedy testy podmíněné a nepodmíněné nezávislosti. Jednou z možností je χ^2 test, který je založen na srovnávání pozorovaných četností

s očekávanými, které jsou počítány za předpokladu nulové hypotézy o nezávislosti těchto proměnných. Při zamítnutí nulové hypotézy zamítáme nezávislost proměnných. Pokud mají testy hladinu významnosti 0,05, tak je předpokládáno, že 1 z 20 zamítnutí je falešná, proto se u velkého množství hypotéz snižuje schopnost postavit tu správnou strukturu. Určení takovýchto nezávislostí je obtížná úloha s mnohdy nedokonalými výsledky. Komplexita algoritmu na nalezení těchto vztahů závisí i na omezení množství rodičů uzlů. [12]

Příklad 1.3.2 *Využijeme dat Zlomenina k nalezení struktury bayesovské sítě pomocí metody omezení vztahů:*

1. *Nejdříve otestujeme nezávislosti dvojic proměnných v datové množině pomocí χ^2 testů:*

   χ^2 test:

   Máme kontingenční tabulku 2 náhodných výběrů X a Y, které nabývají hodnot 0 nebo 1. M je celkový počet pozorování a M[i,j] značí počet pozorování, pro které má X hodnotu i a Y hodnotu j. Testujeme na hladině významnosti $\alpha = 0.05$. Tato kontingenční tabulka je zobrazena níže v tabulce 1.3:

   Nulová hypotéza: Proměnné X a Y jsou nezávislé.

	y_0	y_1	
x_0	M[0,0]	M[0,1]	M[0,y]
x_1	M[1,0]	M[1,1]	M[1,y]
	M[x,0]	M[x,1]	M

Tabulka 1.3: Kontingenční tabulka pro náhodné výběry X a Y

Testová statistika na základě dat $\mathcal{D}$:

$$TS^2_{\chi^2}(\mathcal{D}) = \sum_{i=0}^{r}\sum_{j=0}^{c} \frac{(M[i,j] - \frac{M[x,j]*M[i,y]}{M})^2 * M}{M[x,j] * M[i,y]} =$$

$$= \frac{(M[0,0] \cdot M[1,1] - M[0,1] \cdot M[1,0])^2 \cdot M}{M[0,y] \cdot M[1,y] \cdot M[x,0] \cdot M[x,1]},$$

*která má za platnosti nulové hypotézy o nezávislosti χ^2 rozdělení s parametrem r*c, v tomto případě 1. Nulovou hypotézu o nezávislosti X a Y zamítneme na hladině α, pokud platí:*

$$TS^2_{\chi^2} > \chi^2_{(r\cdot c)}(1-\alpha).$$

V datové množině máma 4 proměnné: Déšť, Obratnost, Uklouznutí a Zlomenina. Pro každou možnou kombinaci provedeme χ^2 test nezávislosti.
Déšť a Uklouznutí:
Testovací statistika:

	u_0	u_1	
d_0	4808	2136	6944
d_1	1061	1995	3056
	5869	4131	10000

Tabulka 1.4: Kontingenční tabulka pro Déšť a Uklouznutí

$$TS^2_{\chi^2}(Zlomenina) = \sum_{i=0}^{1}\sum_{j=0}^{1} \frac{(M[i,j] - \frac{M[x,j]\cdot M[i,y]}{M})^2 \cdot M}{M[x,j]\cdot M[i,y]} =$$

$$= \frac{(4808 - \frac{6944\cdot 5869}{10000})^2 \cdot 10000}{6944 \cdot 5869} + \cdots = 1043.06$$

Dále platí:

$$TS^2_{\chi^2}(Zlomenina) > \chi^2_1(0.95) = 3.84.$$

A jelikož kritická hodnota je menší než testovací statistika, tak nulovou hypotézu o nezávislosti proměnné Déšť a Uklouznutí zamítáme. Což pro nás bude znamenat hranu mezi těmito proměnnými v síti.
Déšť a Obratnost:

Testovací statistika:

$$TS^2_{\chi^2}(Zlomenina) = \sum_{i=0}^{1}\sum_{j=0}^{2} \frac{(M[i,j] - \frac{M[x,j]\cdot M[i,y]}{M})^2 \cdot M}{M[x,j]\cdot M[i,y]} =$$

$$= \frac{(2112 - \frac{6944\cdot 3023}{10000})^2 \cdot 10000}{6944 \cdot 3023} + \cdots = 0.398$$

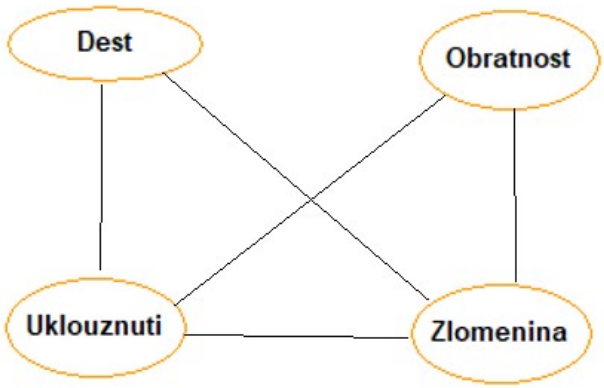
	o_0	o_1	o_2	
d_0	2112	2750	2082	6944
d_1	911	1216	929	3056
	3023	3966	3011	10000

Tabulka 1.5: Kontingenční tabulka pro Déšť a Obratnost

Dále platí:

$$TS^2_{\chi^2}(Zlomenina) < \chi^2_2(0.95) = 5.99.$$

A jelikož kritická hodnota je větší než testovací statistika, tak nulovou hypotézu o nezávislosti proměnné Déšť a Uklouznutí nezamítáme. Což pro nás bude znamenat, že hrana mezi těmito proměnnými v síti nebude.
Dále bychom pokračovali s těmito testy, můžeme využít funkce `chisq.test()` *v Rku. Takhle zjistíme, že nulovou hypotézu nezamítáme pouze v případě dvojice Déšť a Obratnost. Nyní máme představu o naší síti v podobě sítě, která je zobrazena na obrázku 1.7.*



Obrázek 1.7: Bayesovská síť po první fázi testů nezávislosti proměnných

2. *Jako další krok je potřeba otestovat podmíněné nezávislosti, které jsou uvedeny v tabulce 1.6.*

   V tomto případě je možné využít rozšířeného χ^2 testu podmíněné nezávislosti:

$(D \perp Z\|U)$	$(D \perp O\|U)$	$(O \perp Z\|U)$	$(O \perp U\|D)$
$(D \perp U\|Z)$	$(D \perp Z\|O)$	$(O \perp U\|Z)$	$(U \perp Z\|D)$
$(D \perp O\|Z)$	$(D \perp U\|O)$	$(O \perp Z\|D)$	$(U \perp Z\|O)$

Tabulka 1.6: Tabulka potřebných testů podmíněných nezávislostí

Test podmíněné nezávislosti

Budeme zkoumat nezávislost 2 náhodných výběrů X a Y podmíněnou náhodnému výběru Z. M[i,j,k] značí množství pozorování, pro které platí, že hodnota X je i, Y je j a Z je k. $\hat{P}$ značí empirickou pravděpodobnost z dat. Čerpáno z [12].

Nulová hypotéza: Proměnná X a Y jsou nezávislá při známé hodnotě Z.

$$P^*(X, Y, Z) = \hat{P}(Z)\hat{P}(X|Z)\hat{P}(Y|Z)$$

Testovací statistika na základě dat $\mathcal{D}$:

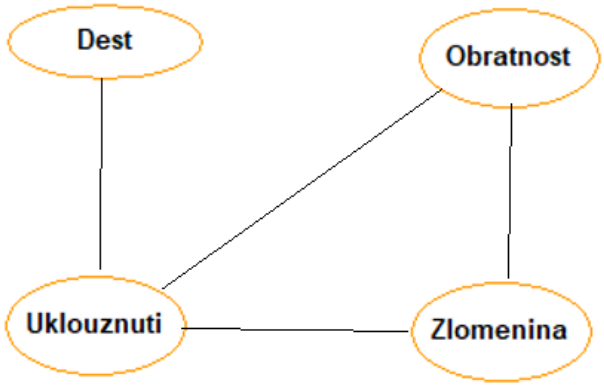
$$TS^2_{\chi^2}(\mathcal{D}) = \sum_{i=0}^{r}\sum_{j=0}^{c}\sum_{k=0}^{s}\frac{(M[i,j,k] - M \cdot \hat{P}(z=k)\hat{P}(x=i|z=k)\hat{P}(y=j|z=k))^2}{M \cdot \hat{P}(z=k)\hat{P}(x=i|z=k)\hat{P}(y=j|z=k)}.$$

Pomocí softwaru R jdou tyto testy provést pomocí funkce `ci.test()`*. Při využití této funkce zjistíme, že nulovou hypotézu nezamítáme jen pro případ $(D \perp Z|U)$. To se promítne do sítě tak, že nám zmizí hrana mezi Deštěm a Zlomeninou, jak je zobrazeno na obrázku 1.8. Jak jde vidět, tak tato struktura je správná, jelikož víme, jak má tato síť vypadat.*

3. *Dále je potřeba stanovit směry hran.*

   (a) *Nejdříve stanovíme množinu proměnných $S_{(r,s)}$ pro každou dvojici proměnných r a s. V této množině budou proměnné, které vybranou dvojici dělají podmíněně nezávislou. V tomto případě vyšla podmíněná nezávislost pouze u dvojice Déšť a Zlomenina na základě Uklouznutí. Takže $S_{(dest,zlomenina)}$ obsahuje Uklouznutí a ostatní množiny $S_{(r,s)}$ jsou prázdné.*

   (b) *Nyní vezmeme všechny dvojice proměnných, které nejsou ve struktuře sousedy, ale mají společného souseda. V tomto případě máme dvo-*



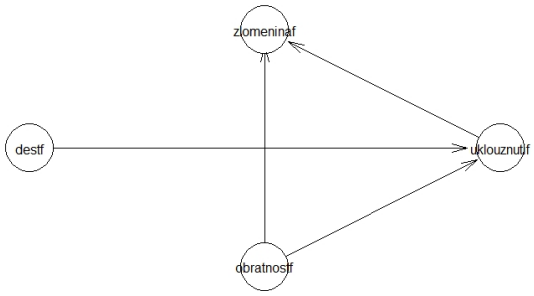
Obrázek 1.8: Bayesovská síť po druhé fázi podmíněného testování nezávislosti proměnných

jice: Déšť a Zlomenina (s Uklouznutím), Déšť a Obratnost (s Uklouznutím). U těchto dvojic zkoumáme, jestli jejich společný soused patří do jejich množiny S. Pokud ne, tak se vytvoří takzvaná V-struktura. U dvojice Déšť a Obratnost jejich společný soused Uklouznutí nepatří do množiny $S_{(dest,obratnost)}$*, jelikož tato množina je prázdná. Takže zde vznikne zmiňovaná V-struktura: Dest* $\rightarrow$ *Uklouznuti* $\leftarrow$ *Obratnost. V druhém případě, u dvojice Déšť a Zlomenina, jejich soused Uklouznutí patří do množiny* $S_{(dest,zlomenina)}$*, takže zde nemá vzniknout V-struktura. A jelikož směr hrany z deště do uklouznutí je už daný, tak směr hrany z uklouznutí do zlomeniny vypadá takto: Uklouznuti* $\rightarrow$ *Zlomenina. Jedině tak zaručíme, že zde nevznikne V-struktura.*

(c) *U zbylých hran zadáme směry tak, aby nevznikl cyklus v grafu. Takže směr hrany mezi Obratností a Zlomeninou musí být následující: Obratnost* $\rightarrow$ *Zlomenina.*

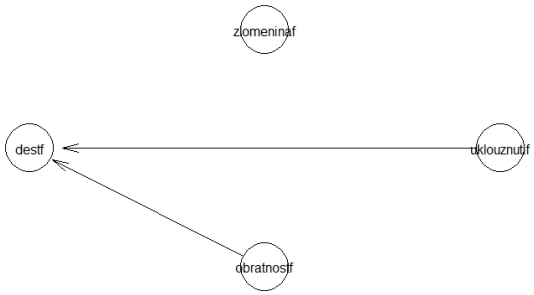
Tímto jsme zpětně vytvořili danou bayesovskou síť, která odpovídá síti ze které jsme data vygenerovali. Tento proces se dá provést funkcí `pc.stable()` *v Rku z balíčku* `bnlearn`*. Výstup této funkce je síť zobrazena na obrázku 1.9.*

Nicméně, musí být řečeno, že bylo použito 10000 pozorování na 4 proměnné, což



Obrázek 1.9: Struktura bayesovské sítě při využití pc.stable funkce pro 10000 pozorování

je velmi velké množství. Při využití náhodně vybraných 100 pozorování z datové množiny Zlomenina je výsledná struktura bayesovské sítě zobrazena na obrázku 1.10, což jak jde vidět není skutečná struktura.



Obrázek 1.10: Struktura bayesovské sítě při využití pc.stable funkce pro 100 pozorování

Určení struktury na základě skórovacích funkcí

Tato forma učení struktury má nyní velkou oblibu a to díky tomu, že jde o optimalizační úlohu. Cílem určení struktury z dat je nalézt DAG $\mathcal{G}$, který maximalizuje $P(\mathcal{G}|\mathcal{D})$, tedy maximalizuje pravděpodobnost vybrané struktury při

daných datech. Nicméně nalezení optimální struktury je NP-obtížný a také NP-úplný problém. [17] Tento problém je náročný také proto, že pokud se rozhodneme o jedné hraně, tak díky této hraně mezi proměnnými mohou být nějaké určité hrany z nebo do těchto proměnných znemožněny, jelikož by nebyla zachována struktura orientovaného acyklického grafu. Tedy vyhledání sítě s maximálním skóre je NP-obtížný problém i při stanovení maximálního počtu 2 rodičů, o čemž hovoří následující věta.

Věta 1.3.5 *Následující problém je NP-obtížný pro jakékoliv $d \geq 2$, $d \in \mathbb{Z}$:*

$$\hat{\mathcal{G}} = \underset{\mathcal{G} \in \boldsymbol{\mathcal{G}}_d}{\operatorname{argmax}} \ score(\mathcal{G}|\mathcal{D}),$$

kde $\boldsymbol{\mathcal{G}}_d = \{\mathcal{G} : \forall i, |Pa^{\mathcal{G}}_{X_i} \leq d\}$ a $score(\mathcal{G}|\mathcal{D})$ je skórovací funkce, která ohodnocuje každého kandidáta DAG $\mathcal{G}$ s respektem k datové množině $\mathcal{D}$.

Pro každou možnou strukturu se vypočítá skóre a hledá se struktura s nejvyšším. Je tedy zvolena skórovací funkce *score* a optimalizační algoritmus, který hledá síť s nejvyšším skóre. Nejvíce využívanými optimalizačními algoritmy jsou Hill-climbing a Tabu hledání, dalšími jsou genetické algoritmy, simulované žíhání a jiné.

- Hill-climbing: je populární zejména kvůli jeho kompromisu mezi výpočetní náročností a kvalitou výstupního modelu. Je to lokální hledání v prostoru grafů. Kardinalita prostoru hledání je super-exponenciální, proto je vhodné při vyšším počtu proměnných tento prostor omezit, nejčastěji počtem rodičů každého uzlu. Algoritmus začíná z inicializačního stavu, kterým může být prázdná, náhodná nebo expertně zadaná struktura. Lokálně postupuje s konečným počtem kroků, kterými mohou být odstranění hrany, přidání hrany, změnění směru hrany. Algoritmus ale vždy vyhodnocuje, zda je daný krok možný, aby graf zachovával podobu DAG. Vždy je vybrán ten krok, který zaznamená největší zvýšení hodnoty skórovací funkce. Algoritmus končí pokud nemá možnost dalšího zlepšení. Jelikož jde o lokální hledání,

tak může lehce skončit u lokálního minima. Využívá se proto mnoho strategií, jak se tomu vyhnout, například restarty a náhodnost. [6]

- Tabu hledaní: tato metoda funguje stejně jako výše uvedený hill-climbing, navíc ale uchovává v paměti seznam nedávných h operací, které byly uplatněny a v dalších krocích nelze použít operace, které by je vyrušily. Takže pokud se algoritmus rozhodne pro vytvoření hrany z $X \to Y$, tak v dalších h krocích tuto hranu nelze odebrat. Také je rozdíl v tom, že se algoritmus neukončí, když nemá možnost zlepšení, ale je stanoven počet kroků, po kterém se algoritmus zastaví, pokud nenalezne lepší strukturu. Výsledkem je potom nejlepší poslední nalezená struktura.

Důležitým krokem mimo výběru algoritmu je i výběr skórovací funkce.

- Věrohodnostní funkce: Věrohodnostní funkce měří pravděpodobnost dat při daném modelu, takže je potřeba najít model, pro který jsou data nejpravděpodobnější, protože funkci věrohodnosti maximalizujeme. Využití věrohodnostní funkce si můžeme ukázat na jednoduchém příkladu. Tento příklad je převzat z [14], strana 101.

  Příklad 1.3.3 *Nechť X_1, X_2, X_3, X_4 je náhodný výběr z alternativního rozdělení s parametrem p, kde $p = 0.2$ nebo $p = 0.8$. Budeme odhadovat parametr, pokud jsme realizací výběru dostali data 1;1;0;1. Pro pravděpodobnost takovýchto výsledků máme:*

  $$P_p(X_1 = X_2 = 1, X_3 = 0, X_4 = 1) = p^3(1 - p).$$

  Tento výsledek pro parametr $p = 0.2$ je 0.0064 a pro $p = 0.8$ je 0.1024. Principem metody maximální věrohodnosti je vzít nejpravděpodobnější výsledek. Pro tento případ tedy parametr $p = 0.8$.

  Věrohodnostní funkci je možné využít jak pro odhad struktury, tak pro odhad parametrů modelu, což bude rozvedeno v další části práce. Při hledání struktury grafu $\mathcal{G}$ se využije maximálně věrohodný odhad parametrů $\hat{\boldsymbol{\theta}}_{\mathcal{G}}$,

který pro kategorická data odpovídá jejich frekvenci v datech. Takové parametry maximalizují funkci věrohodnosti při dané struktuře (viz. kapitola 1.3.4).

Věrohodnostní funkce pro bayesovské sítě při zadané struktuře grafu $\mathcal{G}$ je zadaná následovně: Předpokládejme, že pozorujeme několik IID vzorků z množiny náhodných proměnných $\mathcal{X} = \{X_1, X_2, \ldots, X_k\}$ z neznámého rozdělení $P^*(\mathcal{X})$. Známe možné hodnoty náhodných proměnných. Označme trénovací množinu pozorování jako $\mathcal{D}$, která obsahuje n pozorování z X : $\{x_1, \ldots, x_n\}$, kde x_i pro $i = 1, \ldots, n$ je vektor hodnot proměnných pro dané pozorování, tedy $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$. Pro věrohodnostní funkci platí:

$$\begin{aligned} L(\boldsymbol{\theta}, \mathcal{D}) = \prod_{i=1}^{n} P_{\mathcal{G}}(x_i|\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{j=1}^{k} P(x_{ij}|Pa_{x_{ij}}, \boldsymbol{\theta}) = \\ = \prod_{j=1}^{k}\Big[\prod_{i=1}^{n} P(x_{ij}|Pa_{x_{ij}}, \boldsymbol{\theta})\Big]. \end{aligned} \tag{1.10}$$

Model je dán jako dvojice grafu $\mathcal{G}$ a parametrů $\hat{\boldsymbol{\theta}}_{\mathcal{G}}$. A tedy platí:

$$\max_{\mathcal{G},\boldsymbol{\theta}_{\mathcal{G}}} L(\langle \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}} \rangle, \mathcal{D}) = \max_{\mathcal{G}}[\max_{\boldsymbol{\theta}_{\mathcal{G}}} L(\langle \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}} \rangle, \mathcal{D})] = \max_{\mathcal{G}}[L(\langle \mathcal{G}, \hat{\boldsymbol{\theta}}_{\mathcal{G}} \rangle, \mathcal{D})] \tag{1.11}$$

Tedy k maximalizování věrohodnosti dvojice $(\mathcal{G}, \theta_{\mathcal{G}})$ je nezbytné najít grafovou strukturu $\mathcal{G}$, která má nejvyšší věrohodnost při využití maximálně věrohodného odhadu parametrů. Parametry jsou tedy dané a hledáme strukturu, která maximalizuje funkci věrohodnosti. S každou změnou struktury se nám mění výpočet, protože struktura výpočtu je závislá na struktuře sítě. Jako obvykle se dále pracuje s logaritmickou funkcí věrohodnosti a skórovací funkce metodou maximální věrohodnosti vypadá:

$$score_L(\mathcal{G}|\mathcal{D}) = l(\hat{\boldsymbol{\theta}}_{\mathcal{G}}, \mathcal{D}), \tag{1.12}$$

kde $l(\hat{\boldsymbol{\theta}}_{\mathcal{G}}, \mathcal{D})$ je logaritmus věrohodnostní funkce. Tato funkce se dá vyjádřit také pomocí vzájemné informace.

Definice 1.3.6 *Pro náhodné veličiny X a Y je vzájemná informace mezi X a Y definována jako*

$$I(X,Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} = E\, log \frac{p(X,Y)}{p(X)p(Y)}.[18]$$

Vzájemnou informaci $\mathbf{I}_P(X;Y)$ lze interpretovat jako sílu závislosti mezi X a Y v pravděpodobnosti. Podívejme se na výpočet z našeho příkladu Zlomenina.

Příklad 1.3.4 *Budeme se zabývat pouze proměnnou Déšť a Obratnost. Jako první vypočítáme skóre sítě, kde jsou tyto proměnné nezávislé. Takovou to strukturu označíme jako $\mathcal{G}_0$. Pak platí pro x_i, kde $i = 1, \ldots, n$, že $Pa_{x_i} = \emptyset$. M[y] označuje dané množství pozorování s hodnotou y a parametr $\hat{\theta}$ se rovná empirickému rozdělení pozorovaného v datech, v našem případě frekvenci vybraných dat, například $\hat{\theta}_{d_0} = \hat{P}(d_0) = \frac{M[d_0]}{M}$, kde $M = n$.*

$$score_L(\mathcal{G}_0|Zlomenina) = \prod_{j=\{D,O\}} [\prod_{i=1}^{n} P(x_{ij}|Pa_{x_{ij}}, \boldsymbol{\theta})] =$$

$$= \prod_{j=\{D,O\}} [\prod_{i=1}^{n} P(x_{ij}|\boldsymbol{\theta})] = \prod_{i=1}^{n} P(x_{iD}|\boldsymbol{\theta}) \prod_{i=1}^{n} P(x_{iO}|\boldsymbol{\theta}) =$$

$$= \sum_{i=1}^{n} logP(x_{iD}|\boldsymbol{\theta}) + \sum_{i=1}^{n} logP(x_{iO}|\boldsymbol{\theta}) =$$

$$= \sum_{i=1}^{n} (log\hat{\theta}_D)_i + \sum_{i=1}^{n} (log\hat{\theta}_O)_i = \sum_{s=\{d_0,d_1\}} M[s] log\hat{\theta}_s + \sum_{t=\{o_0,o_1,o_2\}} M[t] log\hat{\theta}_t =$$

$$= M[d_0] log\hat{\theta}_{d_0} + M[d_1] log\hat{\theta}_{d_1} + M[o_0] log\hat{\theta}_{o_0} + M[o_1] log\hat{\theta}_{o_1} + M[o_2] log\hat{\theta}_{o_2}$$

Dané množství se rovnají: $M[d_0] = 6944; M[d_1] = 3056; M[o_0] = 3023; M[o_1] = 3966; M[o_2] = 3011$, *tak po dosazení získáme:*

$$score_L(\mathcal{G}_0|Zlomenina) = 6944 * log(0.6944) + 3056 * log(0.3056) +$$

$$+3023 * log(0.3023) + 3966 * log(0.3966) + 3011 * log(0.3011) =$$

$$= -17053.87560494$$

Nyní stejným způsobem vypočítáme skóre pro síť, která by měla hranu z Deště do Obratnosti, tedy Dest → Obratnost. Tuto síť označíme jako $\mathcal{G}_1$. Platí, že $Pa_{x_{iO}} = x_{iD}$ a $Pa_{x_{iD}} = \emptyset$.

$$score_L(\mathcal{G}_1|Zlomenina) = \prod_{j=D,O}[\prod_{i=1}^{n} P(x_{ij}|Pa_{x_{ij}}, \boldsymbol{\theta})] = \prod_{i=1}^{n} P(x_{iD}|\boldsymbol{\theta}) =$$

$$= \prod_{i=1}^{n} P(x_{iO}|x_{iD}, \boldsymbol{\theta}) = \sum_{i=1}^{n} logP(x_{iD}|\boldsymbol{\theta}) + \sum_{i=1}^{n} logP(x_{iO}|x_{iD}, \boldsymbol{\theta}) =$$

$$= \sum_{i=1}^{n} log(\hat{\theta}_O)_i + \sum_{i=1}^{n} log(\hat{\theta}_{(O|D)})_i =$$

$$= \sum_{s=\{d_0,d_1\}} M[s] log\hat{\theta}_s + \sum_{s=\{d_0,d_1\}} \sum_{t=\{o_0,o_1,o_2\}} M[s,t] log\hat{\theta}_{(t|s)} =$$

$$= M[d_0]log\hat{\theta}_{d_0} + M[d_1]log\hat{\theta}_{d_1} + M[d_0,o_0]log\hat{\theta}_{(o_0|d_0)} + M[d_0,o_1]log\hat{\theta}_{(o_1|d_0)} +$$

$$+ M[d_0,o_2]log\hat{\theta}_{(o_2|d_0)} + M[d_1,o_0]log\hat{\theta}_{(o_0|d_1)} +$$

$$+ M[d_1,o_1]log\hat{\theta}_{(o_1|d_1)} + M[d_1,o_2]log\hat{\theta}_{(o_2|d_1)}$$

Pro dosazení je nutné znát navíc kontingenční tabulku pro Déšť a Obratnost:

	o_0	o_1	o_2	
d_0	2112	2750	2082	6944
d_1	911	1216	929	3056
	3023	3966	3011	10000

Tabulka 1.7: Kontingenční tabulka pro Déšť a Obratnost

Dále musíme vypočítat parametry podmíněné pravděpodobnosti

$$\hat{\theta}_{(O|D)} = \hat{P}(O|D) = \frac{\hat{P}(O,D)}{\hat{P}(O)}$$

Poté výpočet vypadá následovně:

$$score_L(\mathcal{G}_1|Zlomenina) = 6944 * log(0.6944) + 3056 * log(0.3056) +$$
$$+2112 * log(\frac{0.2112}{0.6944}) + 2750 * log(\frac{0.2750}{0.6944}) + \cdots + 929 * log(\frac{0.0929}{0.3056}) =$$
$$= -17053.6765201$$

Tento výsledek je o 0.1990839 *více než v prvním případě a jelikož hodnotu věrohodnostní funkce chceme maximalizovat, tak bychom vybrali síť s hranou mezi Deštěm a Obratností, což ale víme, že není skutečná struktura. Tato hodnota* 0.1990839 *se také rovná hodnotě vzájemné informace vykrácené celkovým množstvím,* $M * I_{\hat{P}}(Dest, Obratnost) = 0.1990839$*. Můžeme si ukázat proč.*

$$score_L(\mathcal{G}_1|Zlomenina) - score_L(\mathcal{G}_0|Zlomenina) =$$
$$= \sum_{i=1}^{n} log(\hat{\theta}_O)_i + \sum_{i=1}^{n} log(\hat{\theta}_{(O|D)})_i - \big[\sum_{i=1}^{n}(log\hat{\theta}_D)_i + \sum_{i=1}^{n}(log\hat{\theta}_O)_i\big] =$$
$$= \sum_{s=\{d_0,d_1\}} \sum_{t=\{o_0,o_1,o_2\}} M[s,t] log\hat{\theta}_{(t|s)} - \sum_{t=\{o_0,o_1,o_2\}} M[t] log\hat{\theta}_t$$

A jelikož $M[x,y] = M \cdot \hat{P}(x,y)$*,* $M[y] = M \cdot \hat{P}(y)$*,* $\hat{\theta}_{(y|x)} = \hat{P}(y|x)$*,* $\hat{\theta}_y = \hat{P}(y)$*, tak po dosazení a upravení dostaneme:*

$$score_L(\mathcal{G}_1|Zlomenina) - score_L(\mathcal{G}_0|Zlomenina) =$$
$$= M \sum_{s=\{d_0,d_1\}} \sum_{t=\{o_0,o_1,o_2\}} \hat{P}(s,t) log\frac{\hat{P}(t|s)}{\hat{P}(t)} =$$
$$= M \sum_{Dest,Obratnost} \hat{P}(Dest, Obratnost) log\frac{\hat{P}(Obratnost|Dest)}{\hat{P}(Obratnost)} =$$
$$= M * I_{\hat{P}}(Dest, Obratnost)$$

Vzájemná informace veličin je nezáporná veličina a $\mathbf{I}_{\hat{P}}(X_i; Pa^{\mathcal{G}}_{X_i}) = 0$*, pokud* $Pa_{X_i} = \emptyset$*. Jelikož se skórovací funkce maximalizuje, tak je možno vidět, že vždy budou preferovány struktury složitější před jednoduššími, což povede k přeučení. To v našem případě znamená, že byla struktura s hranou*

mezi Deštěm a Obratností lepší než struktura bez této hrany, což neodpovídá skutečné struktuře.

Obecně platí:

$$score_L((\mathcal{G}|\mathcal{D}) = M \sum_{i=1}^{n} \mathbf{I}_{\hat{P}}(X_i; Pa_{X_i}^{\mathcal{G}}) - M \sum_{i=1}^{n} \mathbf{H}_{\hat{P}}(X_i). \tag{1.13}$$

Lze si všimnout, že druhá suma nezávisí na struktuře, a proto ji lze ignorovat při porovnávání dvou struktur ze stejné datové množiny. Takže při porovnání struktur ze stejné datové množiny se budou porovnávat jen jejich vzájemné informace. A jelikož je to nezáporná veličina, tak bez zadání omezení ke skórovací funkci, pak věrohodnostní funkce najde vždy strukturu úplnou.

- Bayesovské skóre a BIC: jsou skórovací funkce, které jsou založeny na bayesovském přístupu. V tomto případě je nejistota spojená se strukturou a parametry jí příslušnými. Proto se definuje apriorní rozdělení $P(\mathcal{G})$, které dává apriorní pravděpodobnosti různým strukturám grafu. A také $P(\boldsymbol{\theta}_{\mathcal{G}}|\mathcal{G})$, které dává pravděpodobnosti různým volbám parametrů při daném grafu. Podle Bayesova pravidla pak platí:

$$P(\mathcal{G}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{G})P(\mathcal{G})}{P(\mathcal{D})}, \tag{1.14}$$

kde je jmenovatel normalizační faktor, který při porovnávání struktur zůstává stejný, proto je možné ho vynechat při porovnávání. Poté je bayesovské skóre definováno jako:

$$score_B(\mathcal{G} : \mathcal{D}) = log\, P(\mathcal{D}|\mathcal{G}) + log\, P(\mathcal{G}). \tag{1.15}$$

Díky tomuto přístupu mohou být některé struktury preferovanější před jinými díky stanovení apriorního rozdělení $P(\mathcal{G})$, nicméně, tento člen ve výše uvedeném výrazu je téměř irelevantní v porovnání s prvním výrazem, který

bere v potaz nejistotu s odhadem parametrů. Takže s rostoucím počtem pozorování ztrácí svůj vliv. Při využití řetězového pravidla lze rozložit marginální věrohodnost $P(\mathcal{D}|\mathcal{G})$ na jednu komponentu pro každé lokální rozdělení. Pak pro výraz platí:

$$P(\mathcal{D}|\mathcal{G}) = \int_{\Theta} P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{G}) P(\boldsymbol{\theta}|\mathcal{G}) \mathrm{d}\boldsymbol{\theta} = \prod_{i=1}^{k} \int P(X_i|Pa_{X_i}, \boldsymbol{\theta}_{Xi}) P(\boldsymbol{\theta}_{\boldsymbol{X_i}}|Pa_{X_i}) \mathrm{d}\boldsymbol{\theta}_{X_i} \tag{1.16}$$

kde $P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{G})$ je věrohodnost dat při dané síti $\langle \mathcal{G}, \boldsymbol{\theta} \rangle$ a $P(\boldsymbol{\theta}|\mathcal{G})$ je apriorní rozdělení přes různé hodnoty parametrů pro danou strukturu $\mathcal{G}$.
Tento výraz lze nahradit frekventistickým testem shody jako je BIC (Bayesovské informační kritérium), který je často využíván díky jeho jednoduchosti. Také platí, že BIC konverguje k $log(P(\mathcal{D}|\mathcal{G}))$, pro $n \to \inf$.

$$score_{BIC}(\mathcal{G}|\mathcal{D}) = l(\hat{\theta}, \mathcal{D}) - \frac{log\, M}{2} Dim[\mathcal{G}], \tag{1.17}$$

kde $Dim[\mathcal{G}]$ je dimenze modelu nebo počet nezávislých parametrů v modelu. Jelikož negací tohoto výrazu dostaneme počet bitů potřebných k zakódování modelu, tak je pak tento výraz známý jako minimální deskriptivní délka (MDL). Výraz se dá také zapsat jako:

$$score_{BIC}(\mathcal{G}|\mathcal{D}) = M \sum_{i=1}^{n} \mathbf{I}_{\hat{P}}(X_i; Pa_{X_i}) - M \sum_{i=1}^{n} \mathbf{H}_{\hat{P}}(X_i) - \frac{log\, M}{2} Dim[\mathcal{G}]. \tag{1.18}$$

Při porovnání s věrohodnostní funkcí z předcházejícího bodu jde vidět, že tato funkce penalizuje strukturu za komplexnost. Nicméně, vzájemná informace roste lineárně s M a komplexita logaritmicky. Takže, při vysokém M je kladen důraz na shodu s daty. Asymptoticky BIC preferuje struktury, které přímo odpovídají závislostem v datech.

Příklad 1.3.5 *Nyní pojďme vypočítat BIC skóre struktury Deště a Obratnosti s a bez směrové šipky z minulého příkladu s využitím maximálně věrohodného odhadu parametrů sítě. V minulém příkladu jsme vypočítali,*

že $l_{\mathcal{G}_0}(\boldsymbol{\theta}, Zlomenina) = -17053.8756$ *a* $l_{\mathcal{G}_1}(\boldsymbol{\theta}, Zlomenina) = -17053.6765$. *Proto BIC skóre bude:*

$$score_{BIC}(\mathcal{G}_0|Zlomenina) = l_{\mathcal{G},}(\boldsymbol{\theta}, Zlomenina) - \frac{log(M)}{2} Dim[\mathcal{G}_0] =$$

$$= -17053.8756 - \frac{log(10000)}{2} * 2 = -17063.086$$

$$score_{BIC}(\mathcal{G}_1|Zlomenina) = l_{\mathcal{G}_\infty}(\boldsymbol{\theta}, Zlomenina) - \frac{log(M)}{2} Dim[\mathcal{G}_1] =$$

$$= -17053.6765 - \frac{log(10000)}{2} * 3 = -17067.492$$

Jelikož je skóre vyšší pro strukturu $\mathcal{G}_0$*, tak by výsledná struktura byla bez hran, což odpovídá skutečné struktuře.*

- Akaikeho informační kritérium: další obměnou skórovací funkce může být také známé AIC skóre:

$$score_{AIC}(\mathcal{G}|\mathcal{D}) = l(\hat{\theta}_{\mathcal{G}}|\mathcal{D}) - Dim[\mathcal{G}]. \tag{1.19}$$

Příklad 1.3.6 *Nyní pojdme vypočítat AIC skóre struktury Deště a Obratnosti s a bez směrové šipky z minulého příkladu s využitím maximálně věrohodného odhadu parametrů sítě.*

$$score_{AIC}(\mathcal{G}_0|Zlomenina) = l_{\mathcal{G},}(\boldsymbol{\theta}, Zlomenina) - Dim[\mathcal{G}_0] =$$

$$= -17053.8756 - 4 = -17055.8756$$

$$score_{AIC}(\mathcal{G}_1|Zlomenina) = l_{\mathcal{G}_\infty}(\boldsymbol{\theta}, Zlomenina) - Dim[\mathcal{G}_1] =$$

$$= -17053.6765 - 8 = -17056.6765$$

Jelikož je skóre vyšší pro strukturu $\mathcal{G}_0$*, tak by výsledná struktura byla bez hran, což odpovídá skutečné struktuře.*

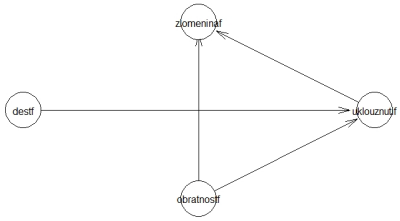
Příklad 1.3.7 *Pomocí softwaru R a balíčku* `bnlearn` *zkusíme najít původní strukturu sítě, ze kterého byla vygenerována data Zlomenina pomocí metody hill-climbing, tabu hledání pro BIC a AIC skórovací funkce.*

- *Pro všechna pozorování (10 000):*

  ```
  bnhcA=hc(zlomeninaD,score="aic")
  plot(bnhcA)
  bntabuA=tabu(zlomeninaD,score="aic")
  bntabuB=tabu(data,score="bic")
  ```

  Nalezená struktura pomocí hill-climbing a AIC skórovací funkcí odpovídá původní skutečné struktuře a je uvedena na obrázku 1.11. Stejně jako při využití tabu hledání s AIC i BIC skórovací funkcí.

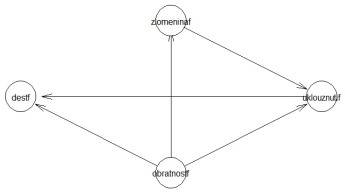
  ```
  bnhcB=hc(zlomeninaD,score="bic")
  ```



Obrázek 1.11: Struktura nalezená pomocí metod: hill-climbing s BIC, tabu s AIC, tabu s BIC

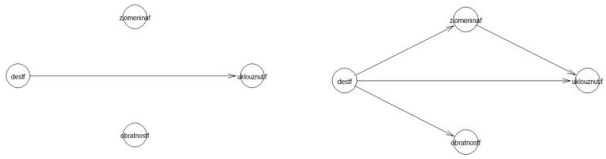
```
plot(bnhcB)
```

Nalezená struktura pomocí hill-climbing a BIC skórovací funkcí neodpovídá původní skutečné struktuře a je uvedena na obrázku 1.12.



Obrázek 1.12: Struktura nalezená pomocí metody hill-climbing a BIC

- *Při využití pouze 100 pozorování:*
  S omezením datové množiny na 100 náhodně vybraných pozorování ani jedna metoda nenašla skutečnou strukturu. Pro AIC i BIC se nalezené struktury rovnaly pro obě optimalizační metody. Obě nalezené struktury jsou uvedeny na obrázku 1.13.



Obrázek 1.13: Nalevo struktura nalezená pro BIC a napravo pro AIC (tabu i hill-climbing shodné výsledky)

Na tomto příkladu jde vidět, že každé metody mohou mít odlišné výsledky. Také můžeme sledovat, že při vyšším množství dat se s větší jistotou dozvíme skutečnou strukturu systému. Nicméně, nelze říci, která metoda je lepší, protože každá metoda může být vhodnější na jiné druhy dat.

1.3.4. Odhad parametrů bayesovské sítě

V minulé části bylo vysvětleno, jak se dá zjistit struktura grafu, dále bude předpokládáno, že je struktura daná a cíl bude odhadnout parametry této sítě. Bayesovská síť je totiž tvořena dvojicí struktury sítě $\mathcal{G}$ s odpovídajícími parametry $\boldsymbol{\theta}_{\mathcal{G}}$. Parametry v diskrétních typech BN odpovídají hodnotám v CPD tabulkách. V realitě je odhad parametrů častěji prováděn z dat, jelikož i zkušení experti nedokáží udat parametry tak detailně, jak je to potřeba například pro složitější CPD tabulky. Jsou zde dva hlavní přístupy a to: odhad založen na metodě maximální věrohodnosti a bayesovský přístup. Čerpáno z [19], [12] a [14].

Metoda maximální věrohodnosti

Metoda maximální věrohodnosti je široce využívaná metoda ve statistice. Tato metoda vychází z frekventistického přístupu. Logaritmická funkce věrohodnosti

pro odhad parametru θ při daných datech $D = \{x_1, \ldots, x_n\}$ je:

$$l(\theta, D) = log\, P(D|\theta) = log \prod_{i=1}^{n} P(x_i|\theta) = \sum_{i=1}^{n} logP(x_i|\theta). \tag{1.20}$$

Věrohodnostní funkce pro bayesovské sítě při zadané struktuře grafu $\mathcal{G}$ je zadaná následovně: Předpokládejme, že pozorujeme několik IID vzorků z množiny náhodných proměnných $\mathcal{X} = \{X_1, X_2, \ldots, X_k\}$ z neznámého rozdělení $P^*(\mathcal{X})$. Známe možné hodnoty náhodných proměnných. Datová množina $\mathcal{D}$ obsahuje n pozorování z $X : \{x_1, \ldots, x_n\}$, kde x_i je vektor hodnot proměnných pro dané pozorování, tedy $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$, pro $i = 1, \ldots, n$. Nechť $Pa^{\mathcal{G}}_{X_j}$ značí rodiče proměnné X_j ve vybrané struktuře $\mathcal{G}$, $Hod(Pa^{\mathcal{G}}_{X_j})$ je množina všech možných kombinací hodnot rodičů uzlu X_j, $Hod(X_j)$ označuje množinu možných hodnot proměnné X_j, $M[x_0, y_4, z_1]$ označuje množství pozorování pro které platí že hodnota $X = x_0, Y = y_4, Z = z_1$, $\hat{\theta}$ označuje empirickou pravděpodobnost. Pak pro logaritmickou funkci věrohodnosti pro diskrétní bayesovskou síť platí:

$$l(\boldsymbol{\theta}, \mathcal{D}) = log \prod_{i} P_{\mathcal{G}}(x_i|\boldsymbol{\theta}) = log \prod_{i} \prod_{j} P(x_{ij}|Pa_{x_{ij}}, \boldsymbol{\theta}) =$$

$$= log \prod_{j} \Big[\prod_{i} P(x_{ij}|Pa_{x_{ij}}, \boldsymbol{\theta}) \Big] = \sum_{j=1}^{k} \Big[\sum_{\mathbf{u}_i \in Hod(Pa^{\mathcal{G}}_{X_j})} \sum_{s_j \in Hod(X_j)} M[s_j, \mathbf{u}_j] log \hat{\theta}_{s_j|\mathbf{u_j}} \Big] \tag{1.21}$$

Pro odhad parametru multinomického rozdělení metodou maximální věrohodnosti podle [19] platí: Nechť $M[q, j, k]$ je počet pozorování v D, pro které platí, že X_i má hodnotu q, a jeden jeho rodič Pa_{X_i} má hodnotu j a druhý hodnotu k. Poté se odhad θ pomocí metody maximální věrohodnosti rovná:

$$\hat{\theta}_{qjk} = \frac{M[q, j, k]}{M[j, k]}. \tag{1.22}$$

Velkým nedostatkem tohoto přístupu je, že není možné odhadnout parametr, pokud $M[j, k] = 0$. Takové případy se s menšími datovými sadami nebo komplikovanějšími strukturami dějí v praxi.

Příklad 1.3.8 *Pojďme se vrátit k příkladu Zlomeniny a vypočítejme parametry u uzlu Uklouznutí s rodiči Déšť a Obratnost. Budeme potřebovat tabulku 1.8 s četnostmi daných kombinací Uklouznutí s rodiči: Déšť a Obratnost.*

	d_0, o_0	d_1, o_0	d_0, o_1	d_1, o_1	d_0, o_2	d_1, o_2
u_0	1246	229	1902	440	1660	392
u_1	866	682	848	776	422	537
$\sum$	2112	911	2750	1216	2082	929

Tabulka 1.8: Tabulka četností pro proměnnou Uklouznutí s rodiči Déšť a Obratnost

Výpočet všech parametrů bude vypadat jako u prvního případu:

$$\hat{\theta}_{(d_0,o_0,u_0)} = \frac{M[d_0, o_0, u_0]}{M[d_0, o_0]} = \frac{1246}{2112} = 0.59$$

Pro srovnání parametr, který byl uveden při generování těchto dat byl 0,6. Takže vidíme správnost odhadu tohoto parametru. Zbylé výsledky odhadů parametrů jsou uvedeny v tabulce 1.9. To stejné bylo provedeno s využitím pouze 100 pozorování, tyto parametry jsou uvedeny v tabulce 1.10. Zde jde vidět, že některé parametry odpovídají skutečnosti, ale některé ne, například pro parametr $\hat{\theta}_{u_0,d_1,o_0}$ vyšla hodnota 0. Takže bychom s jistotou mohli říct, že takový jev nikdy nenastane, což není dobře. Pro srovnání jsou v tabulce 1.11 parametry, které byly využity ke generování dat.

	d_0, o_0	d_1, o_0	d_0, o_1	d_1, o_1	d_0, o_2	d_1, o_2
u_0	0.59	0.25	0.69	0.36	0.8	0.42
u_1	0.41	0.75	0.31	0.64	0.2	0.58

Tabulka 1.9: Tabulka výsledných parametrů pomocí maximálně věrohodného odhadu parametru s využitím 10 000 pozorování

Bayesovská metoda odhadu parametrů

Bayesovský přístup dává možnost přidat do odhadu apriorní rozdělení, které může odhadu přidat znalost uživatele. V odhadu maximální věrohodnosti se ne-

	d_0, o_0	d_1, o_0	d_0, o_1	d_1, o_1	d_0, o_2	d_1, o_2
u_0	0.72	0.00	0.61	0.35	0.84	0.20
u_1	0.28	1.00	0.39	0.65	0.16	0.80

Tabulka 1.10: Tabulka výsledných parametrů pomocí maximálně věrohodného odhadu parametru s využitím 100 pozorování

	d_0, o_0	d_1, o_0	d_0, o_1	d_1, o_1	d_0, o_2	d_1, o_2
u_0	0.60	0.25	0.70	0.35	0.8	0.4
u_1	0.40	0.75	0.30	0.65	0.2	0.6

Tabulka 1.11: Tabulka skutečných parametrů využitých k vygenerování dat

dozvíme, jestli bylo použito 10 nebo milion pozorování, ale při využití bayesovského přístupu lze odhad s málo pozorováními zpřesnit naši apriorní znalostí. Asymptoticky se ale budou rovnat. Je předpokládán obecný problém, který obsahuje datovou množinu $\mathcal{D}$, která obsahuje M nezávislých pozorování z identického rozdělení. Datová množina obsahuje k proměnných z neznámého rozdělení $P^*(X)$. Parametry $\boldsymbol{\theta}$ jsou považovány za náhodnou proměnnou. Takže je potřeba využít pravděpodobnosti k vyjádření počáteční nejistoty spojené s parametry $\boldsymbol{\theta}$ společně s informacemi z dat a využít to v Bayesově pravidlu k vyjádření aposteriorního rozdělení:

$$P(\boldsymbol{\theta}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathcal{D})}. \tag{1.23}$$

První výraz v čitateli je věrohodnostní funkce a druhý výraz je apriorní rozdělení přes všechny možné hodnoty v Θ. Jmenovatel je marginální věrohodnost dat, pro kterou platí:

$$P(\mathcal{D}) = \int_{\Theta} P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{1.24}$$

Tento výraz je normalizační faktor. Při jeho vynechání je aposteriorní rozdělení proporcionální k čitateli. Pokud bychom stanovili rovnoměrné apriorní rozdělení, tak by $P(\theta)$ byla 1 pro jakýkoliv výběr parametru, čímž by se odhad lišil od maximálně věrohodného odhadu jen normalizačním faktorem. Hlavní filozofický rozdíl je ale ve využití aposteriorního rozdělení například pro predikci nového

pozorování.

$$P(x_{n+1}|x_1,\ldots,x_n) = \int P(x_{n+1}|\theta,x_1,\ldots,x_n)\cdot$$

$$\cdot P(\theta|x_1,\ldots,x_n)\mathrm{d}\theta = \int P(x_{n+1}|\theta)P(\theta|x_1,\ldots,x_n)\mathrm{d}\theta$$

Což by pro například pro výpočet hodu mincí s výsledkem 0 nebo 1, znamenalo:

$$P(x_{n+1}=1|x_1,\ldots,x_n) = \frac{1}{P(x_1,\ldots,x_n)}\int \theta\cdot\theta^{M[1]}(1-\theta)^{M[0]}\mathrm{d}\theta$$

Po úpravách by byl výsledek následující:

$$P(x_{n+1}=1|x_1,\ldots,x_n) = \frac{M[1]+1}{M[1]+M[0]+2}$$

Tento odhad je podobný MLE odhadu, až na 1 vzorek navíc ke každé skupině. Čím je M větší, tím má menší sílu, a přibližuje se MLE odhadu. Tomuto odhadu s rovnoměrným apriorním rozdělením se říká Laplaceova korekce. Díky tomu se dá předejít problému, který byl zmiňován u MLE odhadu, a to tomu, že jmenovatel bude 0.

K vyjádření aposteriorního rozdělení pomocí dostatečné statistiky je potřeba, aby podoba apriorního rozdělení byla stejná jako podoba věrohodnostní funkce. Pro multinomické rozdělení, kde prostor parametrů Θ je zadán jako prostor všech nezáporných vektorů $\boldsymbol{\theta} = \langle\theta_1,\ldots,\theta_K\rangle$, takových, že $\sum_k \theta_k = 1$ vypadá věrohodnostní funkce následovně:

$$L(\boldsymbol{\theta},\mathcal{D}) = \prod_k \theta_k^{M[k]}. \tag{1.25}$$

Odpovídajícím apriorním rozdělení pro takovou funkci věrohodnosti je Dirichletovo rozdělení, které zobecňuje Beta rozdělení. Je specifikováno množinou hyperparametrů $\alpha_1,\ldots,\alpha_K$, takových, že platí:

$$\boldsymbol{\theta} \sim Dirichlet(\alpha_1,\ldots,\alpha_K)\ jestliže\ P(\boldsymbol{\theta}) \propto \prod_k \theta_k^{\alpha_k-1}.$$

Při použití Dirichletova apriorního rozdělení je i aposteriorní rozdělení Dirichletové v podobě $Dirichlet(\alpha_1+M[1],\ldots,\alpha_K+M[K])$, kde $M[k]$ značí počet

výskytu hodnoty k. Dirichletovo apriorní rozdělení je konjugované k multinomickému rozdělení. Aposteriorní rozdělení pro predikci nového pozorování pak bude:

$$P(x_{n+1} = k|\mathcal{D}) = \frac{M[k] + \alpha_k}{M + \alpha}.$$

Bayesovský odhad parametrů v BN

Při odhadu parametrů v BN je vycházeno z výše uvedeného postupu.

Definice 1.3.7 (Globální nezávislost parametrů) *Nechť $\mathcal{G}$ je struktura bayesovské sítě s parametry $\boldsymbol{\theta} = (\boldsymbol{\theta}_{X_1|Pa_{X_1}}, \ldots, \boldsymbol{\theta}_{X_n|Pa_{X_n}})$. Apriorní rozdělení $P(\boldsymbol{\theta})$ splňuje globální nezávislost parametrů, pokud má formu:*

$$P(\boldsymbol{\theta}) = \prod_{\mathbf{i}} P(\boldsymbol{\theta}_{X_i|Pa_{X_i}}).$$

Při předpokladu globální nezávislosti parametrů platí:

$$P(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i} P(\boldsymbol{\theta}_{X_i|Pa_{X_i}}|\mathcal{D}). \tag{1.26}$$

A jelikož se dá aposteriorní rozdělení rozdělit do součinu lokálních výrazů, tak se dá tak zjednodušit výraz pro výpočet pravděpodobnosti predikci n+1 pozorování na základě minulých pozorování.

$$\begin{aligned} P(X_{n+1,1}, \ldots, X_{n+1,k}|\mathcal{D}) = \prod_{j} \int P(X_{n+1,j}|Pa_{X_{n+1,j}}, \boldsymbol{\theta}_{X_j|Pa_{X_j}}) \\ P(\boldsymbol{\theta}_{X_j|Pa_{X_j}}|\mathcal{D})\mathrm{d}\boldsymbol{\theta}_{X_j|Pa_{X_j}}. \end{aligned} \tag{1.27}$$

Bayesovský odhad parametrů se tedy řeší pomocí lokálních dekompozicí.

Definice 1.3.8 (Lokální nezávislost parametrů) *Nechť X je proměnná s rodiči $\mathbf{U}$. Řekneme, že apriorní rozdělení $P(\boldsymbol{\theta}_{X|\mathbf{U}})$ splňuje lokální nezávislost parametrů, pokud platí:*

$$P(\boldsymbol{\theta}_{X|\mathbf{U}}) = \prod_{\mathbf{u}} P(\boldsymbol{\theta}_{X|\mathbf{u}}).$$

Pokud jsou podmíněné pravděpodobnostní rozdělení navíc vyjádřena tabulkou (CPT), tak platí:

$$P(\boldsymbol{\theta}|\mathcal{D}) = \prod_j \prod_{Pa_{X_j}} P(\boldsymbol{\theta}_{X_j|Pa_{X_j}}|\mathcal{D}). \tag{1.28}$$

Pokud je apriorní rozdělení $P(\boldsymbol{\theta}_{X|\mathbf{u}})$ Dirichletovo s hyperparametry $(\alpha_{x^1|\mathbf{u}}, \ldots, \alpha_{x^K|\mathbf{u}})$, poté je aposteriorní rozdělení $P(\boldsymbol{\theta}_{X|\mathbf{u}})$ také Dirichletovo s hyperparametry $(\alpha_{x^1|\mathbf{u}} + M[\mathbf{u}, x^1], \ldots, \alpha_{x^K|\mathbf{u}} + M[\mathbf{u}, x^K])$. Každý uzel X_j v bayesovské síti má množinu multinomických rozdělení $\boldsymbol{\theta}_{X_j|Pa_{X_j}}$, jedno pro každou možnou kombinaci hodnot rodičů tohoto uzlu. Každé takové rozdělení bude mít svoje Dirichletova apriorní rozdělení zadané vybranými parametry.

$$\alpha_{X_j|Pa_{X_j}} = (\alpha_{x_j^1|Pa_{X_j}}, \ldots, \alpha_{x_j^{K_j}|Pa_{X_j}})$$

Hyperparametry Dirichletova apriorního rozdělení lze chápat jako imaginární počty pozorování, které byly pozorovány v minulosti, které odráží znalost nebo spíše názor procesu. Pokud nejsou známy žádné apriorní znalosti procesu, tak se volí rovnoměrné rozdělení pro apriorní rozdělení, což je v souladu s maximalizací entropie náhodných proměnných. Poté se volí $\alpha_1 = \alpha_2 = \cdots = \alpha_K$.[5]

Příklad 1.3.9 *Nyní vypočítáme parametry u uzlu Uklouznutí, čemuž jsme se věnovali v předchozím příkladu, ale bayesovským přístupem. Budeme potřebovat stanovit apriorní rozdělení ke každému sloupci CPD tabulky Uklouznutí. Tyto hodnoty budou představovat naši apriorní znalost tohoto procesu a jsou zobrazeny v tabulce 1.12.*

	d_0, o_0	d_1, o_0	d_0, o_1	d_1, o_1	d_0, o_2	d_1, o_2
u_0	30	25	45	10	58	40
u_1	22	60	18	22	15	85
$\sum$	52	85	63	32	73	125

Tabulka 1.12: Tabulka hyperparametrů apriorního rozdělení k výpočtu odhadu parametrů u uzlu Uklouznutí s rodiči Déšť a Obratnost

$$\hat{\theta}_{(d_0,o_0,u_0)} = \frac{\alpha_{u_0|d_o,o_o} + M[d_0, o_0, u_0]}{M[d_0, o_0] + \alpha_{d_0,o}} = \frac{1246 + 30}{2112 + 52} = 0.59$$

Jak jde vidět, tak odhad parametru se po zaokrouhlení nezměnil a to kvůli tomu, že počítáme s 10 000 pozorování. Tím pádem vliv apriorního rozdělení je zanedbatelný. Ostatní výsledky odhadu parametrů jsou v tabulce 1.13. V závorkách vidíme odhad parametru metodou maximální věrohodnosti. Dále se ale podívejme, co to udělá s odhady pro pouze 100 pozorování. Tyto výsledky jsou uvedeny v tabulce 1.14.

	d_0, o_0	d_1, o_0	d_0, o_1	d_1, o_1	d_0, o_2	d_1, o_2
u_0	0.59 (0.59)	0.26 (0.25)	0.69 (0.69)	0.36 (0.36)	0.80 (0.80)	0.41 (0.42)
u_1	0.41 (0.41)	0.74 (0.75)	0.31 (0.31)	0.64 (0.64)	0.20 (0.20)	0.59 (0.58)

Tabulka 1.13: Tabulka výsledných parametrů pomocí bayesovského odhadu parametru s využitím 10 000 pozorování, MLE odhady v závorce

	d_0, o_0	d_1, o_0	d_0, o_1	d_1, o_1	d_0, o_2	d_1, o_2
u_0	0.62 (0.72)	0.28 (0.00)	0.68 (0.61)	0.33 (0.35)	0.80 (0.84)	0.32 (0.20)
u_1	0.38 (0.28)	0.72 (1.00)	0.32 (0.39)	0.67 (0.65)	0.20 (0.16)	0.68 (0.80)

Tabulka 1.14: Tabulka výsledných parametrů pomocí bayesovského odhadu parametru s využitím 100 pozorování, MLE odhady v závorce

Z tabulky 1.14 jde vidět, že bayesovský přístup k odhadu parametrů je zejména vhodný na případy, kdy nemáme dostatečné množství dat, ale máme vcelku dobrou apriorní znalost procesu. Vidíme, že odhad parametru pro parametr $\hat{\theta}_{u_0,d_1,o_0}$ se změnil z 0 na 0,28, což lépe vypovídá o skutečnosti.

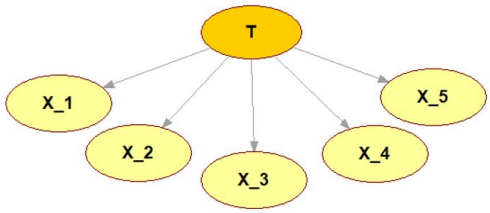
1.3.5. Bayesovská síť vytvořená ke klasifikaci

Klasifikaci spočívá ve vytvoření funkce, která přiřadí třídu vysvětlované proměnné na základě hodnot vysvětlujících proměnných. Tomuto problému se věnuje i strojové učení, ve kterém se jím zabývají rozhodovací stromy, neuronové sítě a jiné. V této části bylo převážně vycházeno z [20] a [12].

Jak bylo zmíněno, tak se bayesovské sítě vytvářejí k dosažení 3 cílů (viz. kapitola 1.3.2) a jedním z nich je právě klasifikace, které se věnuje i praktická část této práce.

Naivní bayesovská síť jako klasifikátor

U naivní bayesovské sítě je stanovena struktura, která je dána klasifikačním uzlem. Tento uzel je poté rodič všech ostatních proměnných a žádné další hrany nejsou povoleny. Takže síť určuje podmíněné pravděpodobnosti každé vysvětlující proměnné na základě třídy vysvětlované proměnné. Klasifikace je provedena na základě Bayesova pravidla, a to tak, že je vypočítaná pravděpodobnost třídy T při daných hodnotách vysvětlujících proměnných $X_1, \ldots, X_k$, a je predikovaná třída s nejvyšší aposteriorní pravděpodobností. Tento model je v jednoduchém provedení zobrazen na obrázku 1.14.



Obrázek 1.14: Ilustrační síť - Naivní bayesovský klasifikátor

Definice 1.3.9 (Naivní bayesovská síť) *Uvažujme strukturu grafu s proměnnými* $\mathbf{U} = \{X_1, \ldots, X_k, T\}$, *kde* $X_1, \ldots, X_k$ *jsou vysvětlujícími proměnnými a* T *je vysvětlovaná proměnná, a kde uzel* T *tvoří kořen sítě, tedy* $Pa_T = \emptyset$, *a každá vysvětlující proměnná má unikátního rodiče, konkrétně* $Pa_{X_i} = \{T\}$, *pro všechny* $1 \leq i \leq k$. *Pro tuto strukturu platí:*

$$P(X_1, \ldots, X_k, T) = P(T) \cdot \prod_{i=1}^{k} P(X_i|T).$$

Z tohoto výrazu lze odvodit, že

$$P(T|X_1, \ldots, X_k) = \alpha \cdot P(T) \cdot \prod_{i=1}^{k} P(X_i|T),$$

kde α *je normalizační konstanta.*

U naivního bayesovského modelu jsou předpokládány podmíněné pravděpodobnostní nezávislosti všech vysvětlujících proměnných X_i na základě vysvětlované proměnné T. Konkrétně, X_1 a X_2 jsou podmíněně nezávislé při daném T, pokud platí: $P(X_1|X_2,T) = P(X_1|T)$ pro všechny možné hodnoty X_1, X_2 *a* T, když $P(T) > 0$. Tento předpoklad je u většiny systémů nerealistický, nicméně, naivní bayesovský klasifikátor (NBK) má překvapivě dobré výsledky. Intuitivně lze předpokládat, že při odlehčení těchto omezení se mohou výsledky zlepšit, takovýmto modelům se budeme věnovat později.

Proč je tedy nutné tvořit klasifikační sítě? Jelikož bayesovské sítě se většinou vytváří na základě skórovací funkce, která měří chybu naučené sítě přes všechny proměnné v modelu. Minimalizací této chyby se ale nezaručí minimalizace lokální chyby v predikování vysvětlované proměnné. Tento problém by se měl podle [20] objevovat ve všech zmíněných skórovacích funkcí. V každé z těchto funkcí se objevuje logaritmická funkce věrohodnosti, pro kterou v případě NBK platí:

$$l(\hat{\boldsymbol{\theta}}_{\mathcal{G}}, \mathcal{D}) = \sum_{i=1}^{n} log\, P_{\hat{\boldsymbol{\theta}}_{\mathcal{G}}}(t_i|x_{i1}, \ldots, x_{ik}) + \sum_{i=1}^{n} log\, P_{\hat{\boldsymbol{\theta}}_{\mathcal{G}}}(x_{i1}, \ldots, x_{ik}), \tag{1.29}$$

kde $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{G}}}$ značí vybrané parametry pro danou strukturu $\mathcal{G}$, data $\mathcal{D} = \{x_1, \ldots, x_n\}$, kde x_i je dán jako vektor hodnot $(x_{i1}, \ldots, x_{ik}, t_i)$ proměnných $X_1, \ldots, X_k, T$. První výraz měří, jak dobře model odhaduje pravděpodobnost klasifikace při daných hodnotách vysvětlujících proměnných. Druhý výraz měří jak dobře síť odhaduje sdruženou pravděpodobnost vysvětlujících proměnných. Jen první výraz tedy určuje kvalitu klasifikace. Bohužel druhý výraz dominuje, pokud je v modelu mnoho vysvětlujících proměnných; jak k roste, pravděpodobnost $X_1, \ldots, X_k$ je menší, poněvadž počet možných kombinací roste exponenciálně s k. Takže lze předpokládat, že výraz $P_{\hat{\boldsymbol{\theta}}}(x_{i1}, \ldots, x_{ik})$ se bude přibližovat nule, a tak výraz $-log\, P_{\hat{\boldsymbol{\theta}}}(x_{i1}, \ldots, x_{ik})$ bude růst více. Ve stejný čas se první výraz víceméně měnit nebude, a tak větší chyba v klasifikaci nebude reflektována ve skórovací funkci. Takže využití skórovacích funkcí k vytvoření bayesovské sítě pro klasifikaci může způsobit špatné klasifikační výsledky. Proto se ke klasifikaci využívají sítě jako je NBK a jemu podobné, které jsou vytvářeny s omezením pro strukturu sítě k

zajištění lepších klasifikačních výsledků.

K potvrzení výsledků [20] vytvořili experiment, kde byla srovnávána klasifikační přesnost pro NBK a pro neomezené sítě (sítě bez omezení na vytváření struktury) se skórovací funkcí MDL na 25 různých datových množinách z UCI repositáře. Výsledky ukázaly, že pro 6 modelů NBK měl signifikantně lepší výsledky než model vytvořený na základě MDL skóre a pro 6 modelů byly výsledky signifikantně horší. Při zkoumání, u kterých z těchto datových množin byly výsledky pro MDL skóre horší se ukázalo, že obsahovaly více než 15 proměnných, a také výsledná síť měla málo proměnných v Markov blanket vysvětlované proměnné. Například datová množina o 36 proměnných měla v Markov blanket u vysvětlované proměnné pouze 5 proměnných, a jelikož při známých hodnotách proměnných v Markov blanket je proměnná nezávislá na ostatních proměnných v síti, tak klasifikaci určovalo pouze těchto 5 hodnot. Takže model sám provádí selekci proměnných, což může být u některých modelů prospěšné, ale u komplikovanějších problémů to může způsobit velmi zhoršenou klasifikaci. Tento problém se vyskytuje i v datech v praktické části této práce.

Řešením tohoto problému by bylo omezit logaritmickou funkci věrohodnosti pouze na první výraz, tedy

$$l_c(\hat{\boldsymbol{\theta}}_{\mathcal{G}}, \mathcal{D}) = \sum_{i=1}^{n} log\, P_{\hat{\boldsymbol{\theta}}_{\mathcal{G}}}(t_i | x_{i1}, \ldots, x_{ik}). \tag{1.30}$$

Pro tento výraz by už ale neplatilo, že odhad parametrů je maximalizován při využití MLE odhadů pro danou strukturu, takže by se tato funkce musela maximalizovat přes prostor všech možných parametrů pro každou kandidátní strukturu. Tento problém se pak stává velmi výpočetně náročný, a tak jeho řešení podle [20] zůstává otevřenou otázkou.

Rozšířená naivní bayesovská síť

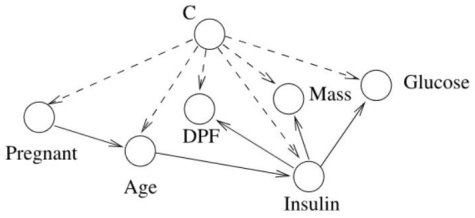
Jak bylo zmíněno, tak předpoklady naivní bayesovské sítě o podmíněné nezávislosti vysvětlujících proměnných jsou většinou nerealistické, a proto by odlehčením těchto předpokladů mohl vzniknout model s lepšími výsledky. Řešením

je tedy tato omezení vynechat a naopak vynutit strukturu naivního bayese, tedy vynutit šipky z vysvětlované proměnné do všech vysvětlujících proměnných a pak dále hledat síť, která bude maximalizovat skórovací funkci. Vynucením šipek z vysvětlované proměnné se zajistí, že informace z každé proměnné bude zahrnuta do predikce. Tedy všechny proměnné budou obsažené v Markov blanket klasifikačního uzlu. Nicméně, tato úloha se stává stejně výpočetně obtížnou jako hledání struktury obecné bayesovské sítě, takže není zaručené, že lze nalézt globální řešení.

V takové struktuře hrana mezi X_i a X_j implikuje, že vliv X_i na T závisí také na X_j, což je aplikováno na následující příklad, který je převzat z [20], str.140.

Příklad 1.3.10 .

Na obrázku 1.15 je zobrazen speciální typ rozšířené naivní bayesovské sítě. Vliv proměnné glukóza na klasifikační proměnnou C závisí na hodnotě inzulínu. U NBK modelu by vlivy proměnných na klasifikační uzel byly nezávislé mezi sebou. Pokud by hodnota glukózy byla nepravděpodobná (v tomto případě by $P(g|c)$ byla nízká), tak by byla nepřekvapivá také nepravděpodobná hodnota inzulínu, což je její korelovaná proměnná (v tomto případě by to znamenalo, že $P(g|c,i)$ je vysoká). V takové situaci by NBK přepenalizovala pravděpodobnost hodnoty klasifikačního uzlu s ohledem na 2 nepravděpodobné hodnoty, kdežto rozšířený NBK by to neudělal.



Obrázek 1.15: TAN model z datového setu "pima"[20]

Stromově rozšířená naivní bayesovská síť

Kouzlo tohoto modelu spočívá v tom, že by měl mít vylepšené klasifikační výsledky oproti modelu naivního bayese, dále by neměl mít nerealistické předpoklady o nezávislosti proměnných, a jeho sestavení pro optimální řešení by měl být problém s řešením v polynomiálním čase. Takzvaný tree-augmented naive bayesian network model, dále jako TAN model, je zadaný tak, že klasifikační uzel nemá žádné rodiče a každá jiná proměnná má za rodiče klasifikační uzel a maximálně jednu jinou proměnnou. Síť uvedená v předchozím příkladě na obrázku 1.15 má takovou strukturu. Procedura na učení takové struktury je založena na známé metodě podle [21].

Procedura k vytvoření TAN modelu je složena z 5 kroků:

1. Vypočítej $I_{\hat{P}_D}(X_i; X_j|T)$ mezi všemi vysvětlujícími proměnnými, $i \neq j$, kde:

$$I_P(\mathbf{X}; \mathbf{Y}|\mathbf{Z}) = \sum_{\mathbf{x},\mathbf{y},\mathbf{z}} P(\mathbf{x}, \mathbf{y}, \mathbf{z}) \; log \; \frac{P(\mathbf{x}, \mathbf{y}|\mathbf{z})}{P(\mathbf{x}|\mathbf{z})P(\mathbf{y}|\mathbf{z})}$$

   je funkce podmíněné vzájemné informace. Zhruba řečeno je tato funkce měřítkem, jak moc informace $\mathbf{Y}$ poskytuje ohledně $\mathbf{X}$ při známé hodnotě $\mathbf{Z}$.

2. Postav kompletní neorientovaný graf, kde jsou uzly proměnné z $\mathcal{X}$. Váhy mezi X_i a X_j odpovídají hodnotě $I_{\hat{P}_D}(X_i; X_j|T)$.

3. Postav maximální stromovou kostru.

4. Transformuj výsledný neorientovaný strom vybráním proměnné, která bude kořenem stromu a nastavením směrů všech hran z této proměnné.

5. Zkonstruuj TAN model přidáním klasifikačního uzlu T a vynucením hran vedoucích z uzlu T do všech vysvětlujících proměnných X_i. [20]

V tomto případě se logaritmická funkce věrohodnosti transformuje do této podoby:

$$l(\hat{\boldsymbol{\theta}}^{\boldsymbol{T}}_{\boldsymbol{\mathcal{G}}}, \mathcal{D}) = N \cdot \sum_{X_i} I_{\hat{P}_D}(X_i; Pa_{X_i}) + konstatni \;\; vyraz, \tag{1.31}$$

takže maximalizování této funkce je rovno maximalizování $\sum_{X_j} I_{\hat{P}_D}(X_j; Pa_{X_j})$, což má podle [20] komplexitu $\mathcal{O}(n^2 \cdot N)$.

Kapitola 2

Praktická část

Tato část práce se bude věnovat zpracování vybraných dat a vytvoření modelu logistické regrese, PRIMA-PI a bayesovské sítě ke stanovení prognózy pacientů s folikulárním lymfomem. Naše cílená proměnná u které budeme provádět predikci a klasifikaci bude EFS24. Souhrn všech proměnných se kterými pracujeme je uveden na začátku této práce v tabulce 1. Všechny metody a jejich výsledky budou porovnány k porozumění přínosů a nedostatků bayesovských sítí.

2.1. Data

Studie zahrnuje 1401 pacientů s folikulárním lymfomem z registru České Lymfatické studijní skupiny, diagnostikovaných mezi 10. 4. 2000 a 28. 12. 2016. Cílem je predikovat hodnoty EFS24, která nabývá binárních hodnot. Tato proměnná určuje buď, zda pacient zemřel, nebo, zda se objevil relaps nebo progrese nemoci do 2 let od první diagnózy.

2.1.1. Úprava dat

Základní soubor dat obsahoval 1401 pozorování o 40 proměnných. K další analýze byly použity proměnné stanovené při diagnóze: věk, pohlaví, stupeň lymfomu, uzlinové lokalizace, kostní dřeň, postižené extra nodální lokalizace, velikost tumoru, celkové příznaky, klinické stádium, performance dle ECOG, LDH vyšší než norma, B2m, leukocyty, lymfocyty, hemoglobin, trombocyty, chemoterapie.

Jelikož u některých těchto proměnných byly záznamy neúplné, nebo neodpovídaly zadanému formátu, tak se soubor dat musel dále upravit. Úpravy byly provedeny v softwaru R.

2.1.2. Neúplné záznamy

Následující tabulka uvádí proměnné, které měly chybějící hodnoty a jejich počet. Tato pozorování byla odstraněna, tedy celé řádky. Proměnná ASCT nakonec nebyla použita. Alternativou k tomuto postupu by bylo vytvořit novou kategorii v každé proměnné, to by ale způsobovalo problémy v algoritmech kvůli malému množství vzorků při fragmentaci dat. Tato strategie byla uplatněna pouze u proměnné Největší tumor, kde byla vytvořena nová proměnná Není známo. Další možností by byla vhodná imputace dat. Druhá tabulka uvádí spojité proměnné, u kterých byly hodnoty nahrazeny jejich mediánem. Medián byl zvolen, protože proměnné někdy obsahovaly odlehlá pozorování a průměr byl jimi ovlivněn.

Proměnná	**Počet odstraněných záznamů**
Stupeň lymfomu	1
Uzlinové lokalizace	1
Kostní dřeň	24
Celkové příznaky	7
Performance dle ECOG	8
LDH vyssi nez norma	9
Chemoterapie	3
ASCT	1

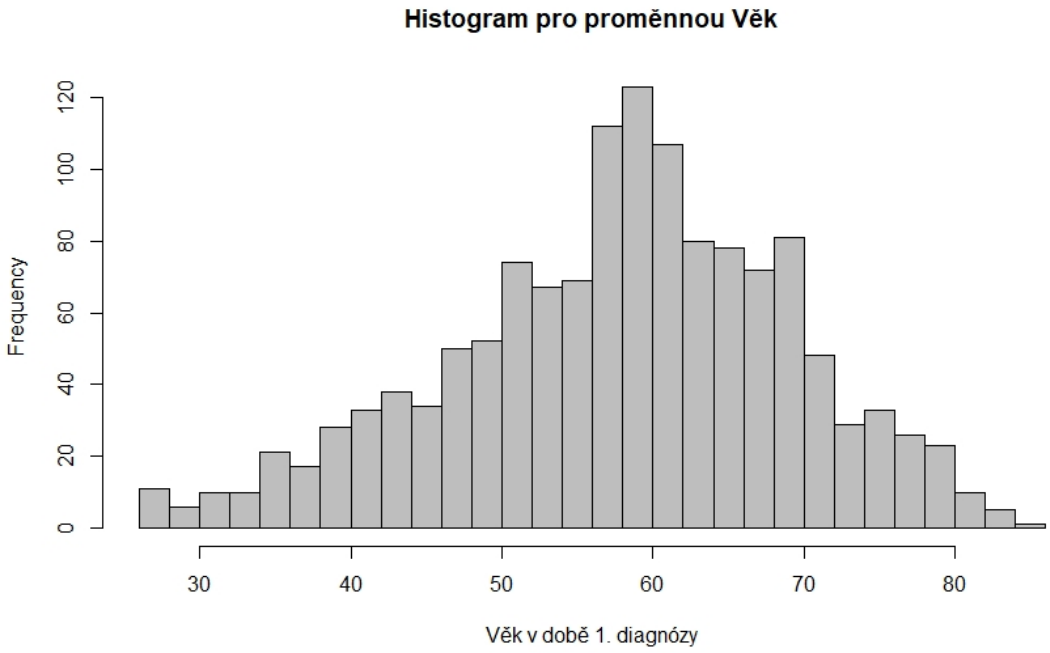
Tabulka 2.1: Počet odstraněných záznamů podle proměnné

2.1.3. Popis dat

Upravená data obsahovala celkem 1348 pacientů, kde 545 pozorování byli muži a 803 ženy. Průměrný věk všech pacientů byl 58,3 let. Histogram proměnné věk je uveden níže na obrázku 2.1. U této proměnné byl proveden Shapiro-Wilkův test normality, který nulovou hypotézu o normalitě zamítl.

Proměnná	Počet nahrazených hodnot	Medián
B2m	147	2.40
Leukocyty	12	6.61
Lymfocyty	27	1.45
Hemoglobin	3	135.00
Trombocyty	13	329.00

Tabulka 2.2: Počet a hodnota nahrazených hodnot u neúplných záznamů u spojitých proměnných



Obrázek 2.1: Histogram pro proměnnou Věk v době první diagnózy

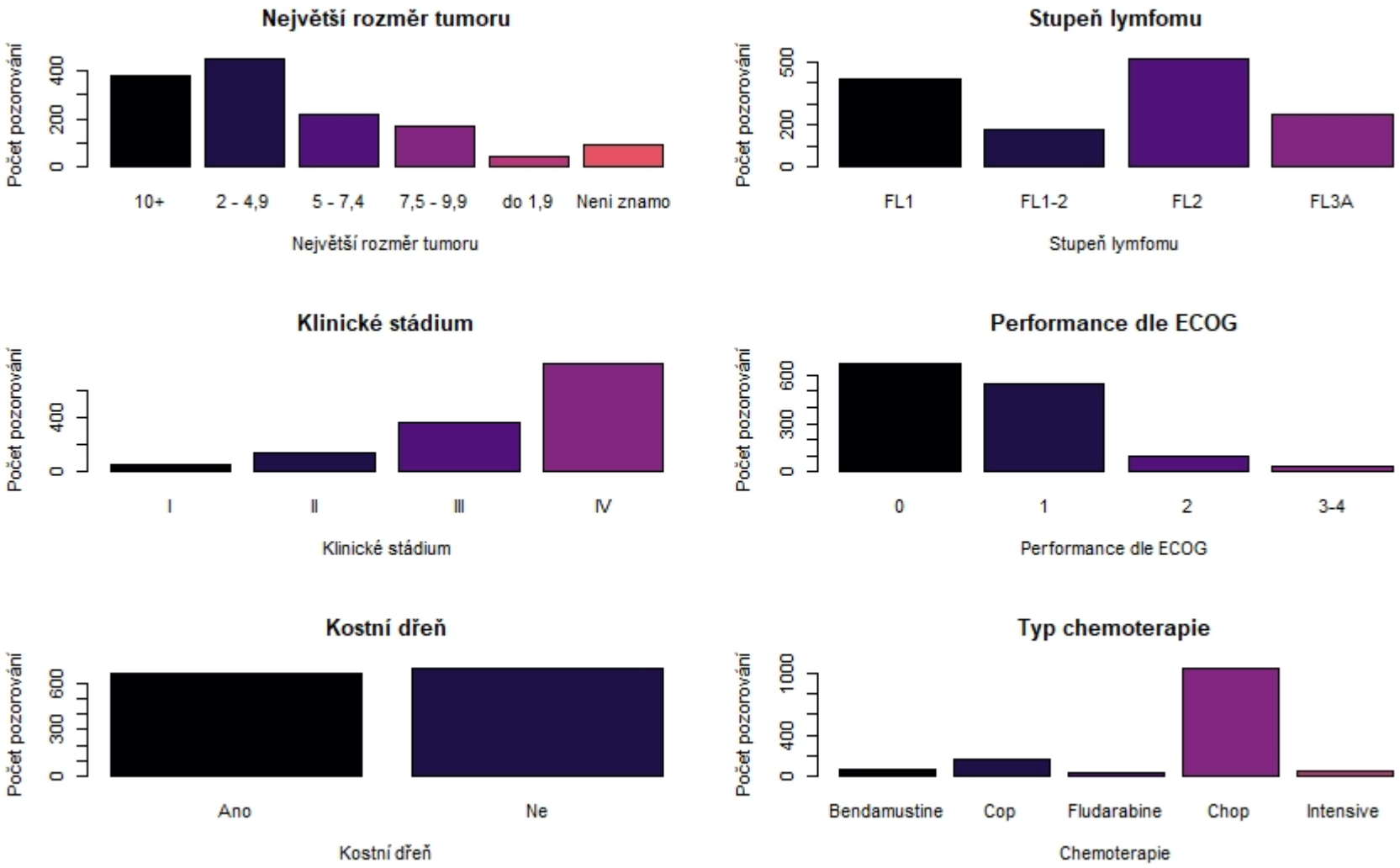
Proměnná Uzlinové lokalizace nabývala diskrétních hodnot od 0 do 11 s průměrem 5,4. Postižené extra nodální lokalizace nabývaly diskrétních hodnot od 0 do 8 s průměrem 0,94. Postižení kostní dřeně vykazovalo 659 pacientů oproti 689 pacientů bez postižení. Stupeň lymfomu byl nejvíce zastoupen FL 1 a FL 2, což je zobrazeno na obrázku 2.2 spolu s rozdělením kategorických proměnných: Největší rozměr tumoru, Klinické stádium, Performance dle ECOG, Kostní dřeň a Typ chemoterapie. Celkové příznaky byly zaznamenány u 421 pacientů oproti

927 pacientů bez příznaků. LDH norma byla překročena u 588 pacientů. Nejvíce pacientů dostalo typ chemoterapie Chop a to 1039 pacientů. Číselné charakteristiky u spojitých proměnných jsou uvedeny v tabulce 2.3. EFS24, která udává událost progrese, relapse nebo smrti do 24 měsíců, se vyskytla u 253 (19 %) pacientů a u 1095 (81 %) se nevyskytla.

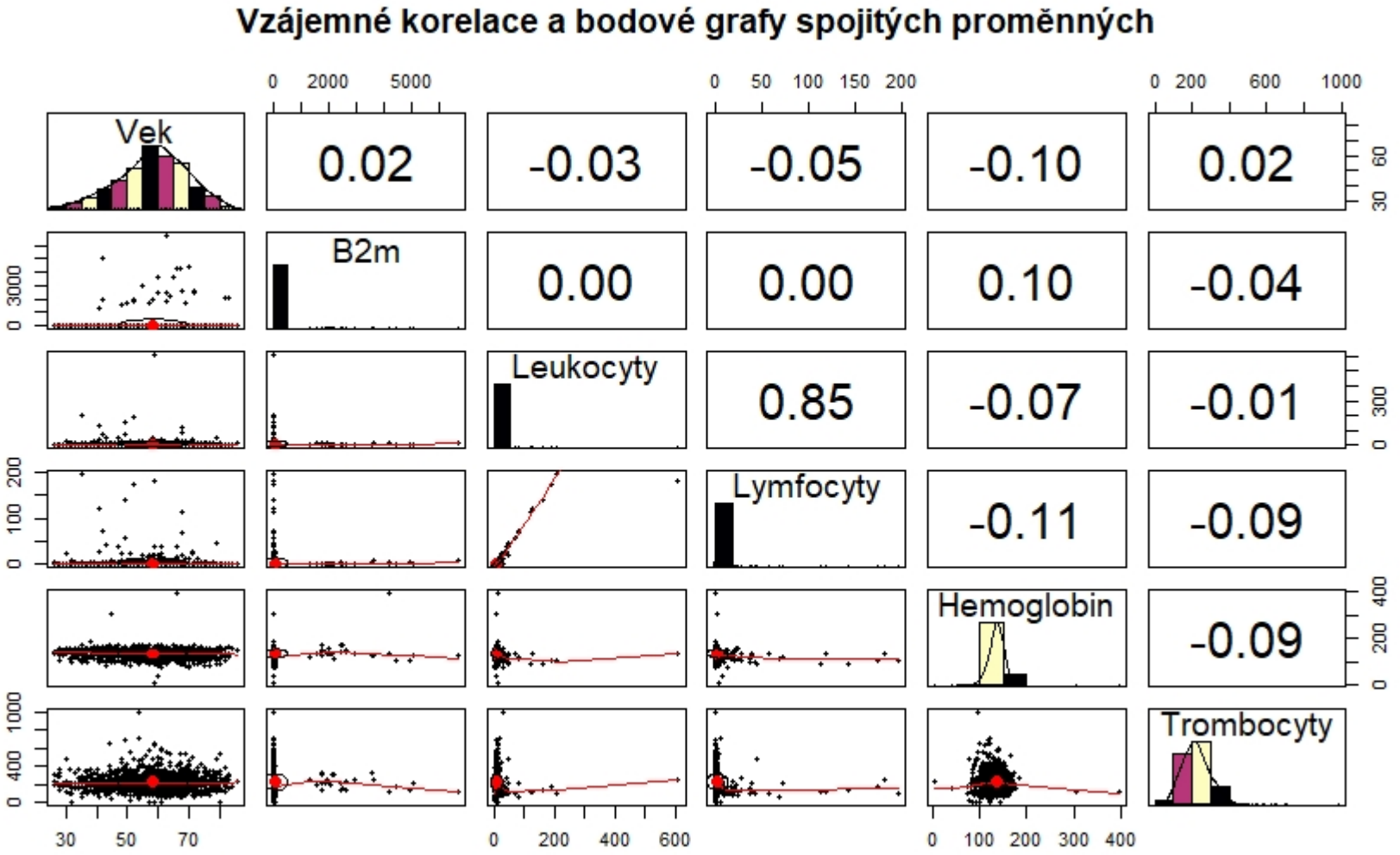
U spojitých proměnných byly vykresleny jejich bodové grafy a vypočítány korelace, jak je uvedeno na obrázku 2.3. Jak jde vidět, tak lymfocyty a leukocyty jsou silně pozitivně provázány. Tyto proměnné nebyly použity spolu v modelu logistické regrese, protože pak by vznikla multikolinearita. U ostatních nebyl detekován žádný významnější vztah.

Čís. char.	**Věk**	**B2m**	**Leukocyty**	**Lymfoc.**	**Hemoglobin**	**Tromb.**
Minimum	26.00	0.17	0.50	0.00	4.59	1.45
D. kvartil	51.00	1.88	5.40	0.99	125.00	169.00
Medián	59.00	2.40	6.61	1.45	135.00	218.00
Průměr	58.29	59.48	8.75	2.96	134.20	228.23
H. kvartil	66.00	3.30	8.43	2.05	146.00	274.00
Maximum	86.00	6710.00	608.00	196.53	395.00	988.00

Tabulka 2.3: Číselné charakteristiky spojitých proměnných



Obrázek 2.2: Rozdělení dat podle vybrané proměnné



Obrázek 2.3: Vzájemné korelace a bodové grafy spojitých proměnných

2.2. Logistický regresní model

Cílem této práce je porovnat klasické metody využívané k predikci prognózy s bayesovskými sítěmi. Taková metoda je logistická regrese, která modeluje střední hodnotu dichotomické závislé proměnné, která zde odpovídá proměnné EFS24. Více k teorii v kapitole 1.2.1. K vytvoření modelů byla data rozdělena na trénovací a testovací sety v poměru 3:1, abychom měli představu, jak bude model reagovat na data, která neviděl.

```
smp_size <- floor(0.75 * nrow(data24))
set.seed(123)
train_ind <- sample(seq_len(nrow(data24)), size = smp_size)
trainingData <- data24[train_ind, ]
testData <- data24[-train_ind, ]
```

Z trénovacích dat byl vytvořen model logistické regrese. Modely byly vytvořeny v softwaru R funkcí glm. Nejprve byl postaven model se všemi proměnnými, které byly vysvětleny v tabulce 1. Poté pomocí funkce step a také pomocí různých pokusů založených na p-hodnotě proměnných a v závislosti na hodnotě AUC (Area under the curve - plocha pod křivkou) byl vybrán model, který měl nejvyšší hodnotu AUC na testovacích datech. Model se všemi proměnnými dosahoval hodnoty AUC 0,678.

2.2.1. Výsledky modelu logistické regrese

Výsledný model byl využit na predikci EFS24 pro pacienty z testovací množiny, která měla 337 pozorování. Vypadal následovně:

```
   glm(formula = EFS24 ~ PerformancedleECOGWHO + Pohlavi +
Kostnidren + LDHvyssineznorma + Hemoglobin + Trombocyty Vek + Chemo
+ B2m, family = "binomial", data = trainingData)
```

Tabulka 2.4 popisuje výsledný model a jeho parametry u každé proměnné a jejich směrodatné chyby v závorce. Podle hvězdiček u parametrů proměnných je zřejmé, že většina proměnných je statisticky významná alespoň na 10% hladině významnosti.

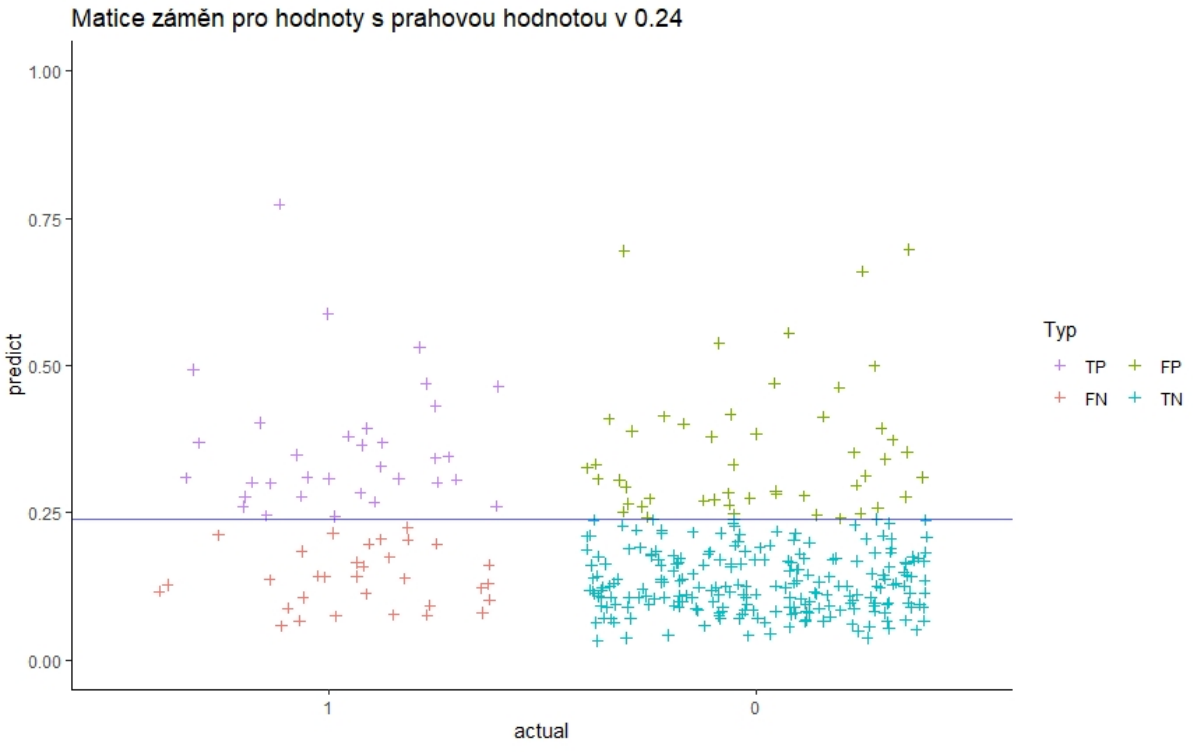
	Dependent variable:
	EFS24
PerformancedleECOGWHO1	0.211 (0.193)
PerformancedleECOGWHO2	0.785^{***} (0.300)
PerformancedleECOGWHO3-4	1.086^{**} (0.488)
Pohlavizena	-0.431^{**} (0.180)
KostnidrenNe	-0.546^{***} (0.187)
LDHvyssineznorma1	0.438^{**} (0.176)
Hemoglobin	-0.018^{***} (0.005)
Trombocyty	-0.002^{*} (0.001)
Vek	0.017^{**} (0.008)
ChemoCop	0.860^{*} (0.485)
ChemoFludarabine	1.558^{**} (0.629)
ChemoChop	0.346 (0.449)
ChemoIntensive	−0.151 (0.704)
B2m	0.0002 (0.0002)
Constant	−0.112 (0.968)
Observations	1,011
Log Likelihood	−434.078
Akaike Inf. Crit.	898.157
Note:	*p<0.1; **p<0.05; ***p<0.01

Tabulka 2.4: Souhrn nejlepšího logistického regresního modelu: hodnoty v druhém sloupci odpovídají hodnotě odhadu parametru pro vybranou proměnnou a hodnoty v závorce značí směrodatnou chybu

Model dosahoval hodnoty AUC 0,689. Při stanovení prahové hodnoty 0,24 jsou výsledky společně se statistikami uvedeny v tabulce 2.5. Predikované hodnoty jsou graficky zakresleny na obrázku 2.4. Z obrázku 2.4 jde vidět, že model není dokonalý, ale dosahuje alespoň vcelku dobré specificity, tedy dokáže identifikovat většinu pozorování, u kterých událost nenastala.

Skutečné hodnoty	Predikce		Statistika	
	0	1	Senzitivita	0.51
0	222	49	Specificita	0.82
1	32	34	F1	0.46

Tabulka 2.5: Matice záměn pro model logistické regrese při zvoleném prahu 0,24



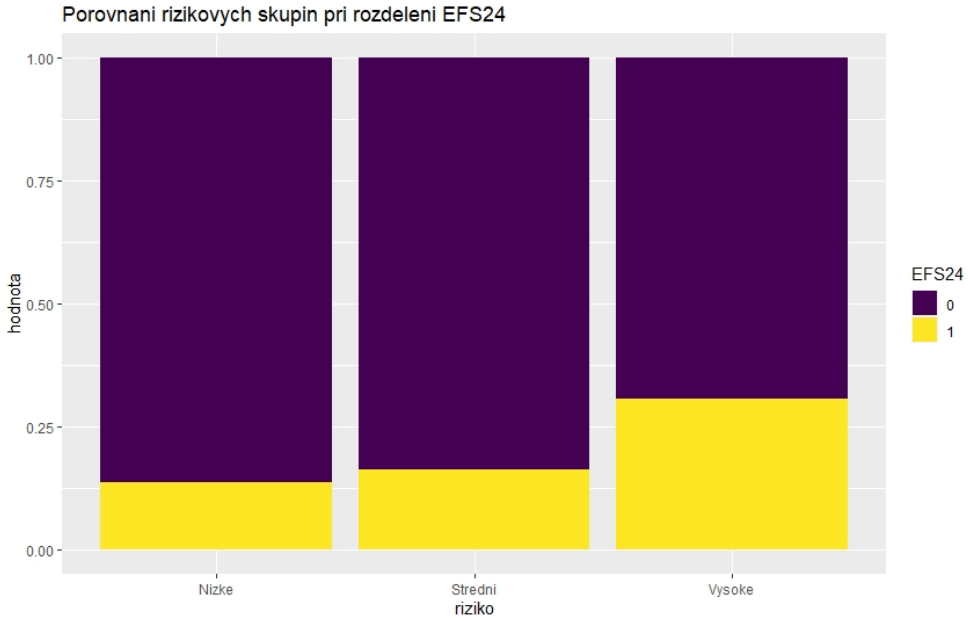
Obrázek 2.4: Grafické zobrazení predikovaných hodnot klasifikace

2.3. PRIMA-PI

Pro vypočítání PRIMA prognostického indexu na daných datech bylo potřeba vytvořit novou proměnnou, která na základě hodnoty B2m a kostní dřeně, určila kategorii rizika každého pozorování. Postup z Rka je zde vypsán:

```
dataB3$PI2[data24$B2m>3]="High"
dataB3$PI2[data24$B2m<=3 & data24$Kostnidren=="Ano"]="Medium"
dataB3$PI2[data24$B2m<=3 & data24$Kostnidren=="Ne"]="Low"
```

Poté stačilo vypočítat kontingenční tabulku v procentech pro proměnné Kategorie rizika a EFS24, a tu graficky zobrazit, viz obrázek 2.5. PRIMA-PI rozděluje data vcelku rovnoměrně do těchto 3 skupin, a to tak, že podíl kategorií střední a vysoké riziko má po aplikování PRIMA-PI 30 % z celkového počtu pozorování a nízké riziko má 40 % z celkového počtu pozorování. Jelikož ale proměnná EFS24 není symetricky rozdělená, u všech skupin nabývá EFS24 hodnoty 1 pouze pro méně než 30 % z celkového množství pozorování ve vybrané skupině. Nicméně, výsledky odpovídají předpokladu, že pacienti s nejvyšším rizikem budou mít nejvyšší výskyt EFS24.



Obrázek 2.5: Grafické zobrazení výsledků při použití PRIMA-PI: proporcionální rozdělení EFS24 v kategoriích rizika

2.4. Bayesovské sítě

V této části se dostáváme ke zpracování dat k vytvoření modelu bayesovské sítě. K vytvoření takového modelu byl využit software GeNIe, který nyní představíme.

Software GeNIe

GeNIe (odvozeno od Graphical Network Interface) Modelátor je využíván k vytváření grafických rozhodovacích modelů. Byl vytvořen Laboratoří rozhodovacích systémů na University of Pittsburgh mezi lety 1995 a 2015. Původně byl vytvořen hlavně ke vzdělávání a výzkumu, díky jeho oblíbenosti byl využíván také státními aparáty a komerčními uživateli. Díky tomu vznikla firma BayesFusion, LLC, která prodává licenci k tomuto programu a poskytuje ho zdarma k využití na vzdělávání a akademický výzkum.

K vytvoření modelu bayesovské sítě si nejdříve musíme upravit data do cílené formy. A jelikož jsme si ověřili, že žádná ze spojitých proměnných nemá normální rozdělení, tak jsme zvolili diskrétní typ bayesovských sítí. Nejdříve jsme si vytvořili kategorické proměnné ze spojitých a diskrétních proměnných. Byla použita funkce `dicretize` z balíčku bnlearn v softwaru R. Vhodným řešením pro praktické využití je také rozdělení do těchto skupin za pomoci experta. My jsme ale rozdělili proměnné do 4 kategorií, které jsou zobrazeny v tabulce 2.6 za pomoci zmiňované funkce. Dále jsme opět pracovali s rozdělenými daty na trénovací a testovací datovou množinu.

	Proměnná					
	Věk	B2m	Leukoc.	Lymfoc.	Hemoglo.	Tromboc.
Intervaly	[26,51]	[0.17,1.88]	[0.5,5.4]	[0,0.99]	[4.59,125]	[1.45,169]
	(51,59]	(1.88,2.4]	(5.4,6.61]	(0.99,1.45]	(125,135]]	(169,218]
	(59,66]	(2.4,3.3]	(6.61,8.43]	(1.45,2.05]	(135,146]	(218,274]
	(66,86]	(3.3,6710]	(8.43,608]	(2.05,197]	(146,395]	(274,988]

Tabulka 2.6: Kategorizace spojitých proměnných

Nahráli jsme si trénovací datovou množinu upravených dat a zkoušeli najít síť s nejlepšími predikčními výsledky. Při využití bayesovských sítí jako klasifikátorů

jsou nejvhodnější tyto typy: naivní bayesovská síť, rozšířená naivní bayesovská síť a stromově rozšířená naivní bayesovská síť (viz. kapitola 1.3.5).

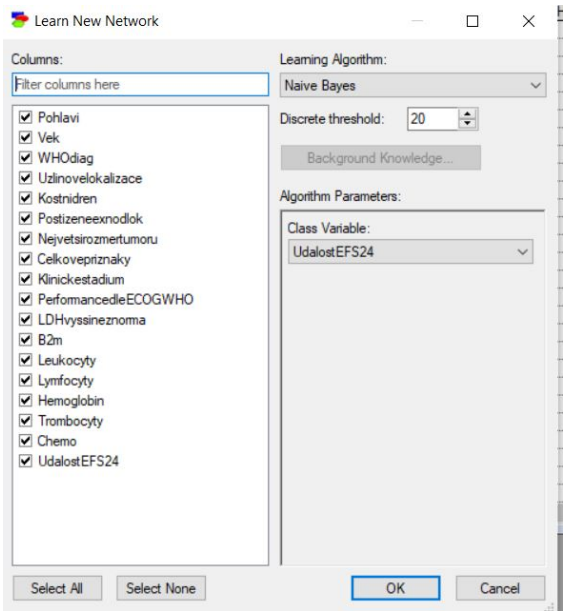
2.4.1. Naivní bayesovská síť

Naivní bayesovská struktura sítě předpokládá nezávislost vysvětlujících proměnných, což se zobrazí v tom, že ve struktuře jsou zakázány hrany mezi vysvětlujícími proměnnými. V tomto příkladě tento předpoklad určitě není splněný. Nicméně tyto modely vykazují i přes tento silný předpoklad dobré výsledky. Struktura u těchto sítí je dána. Tento model je založen na vynucení přímých vztahů, tedy hran, z vysvětlované proměnné do vysvětlujících proměnných. Pokud by hrany vedly opačným směrem, tak by byla výsledkem obrovská CPD tabulka, která by měla všechny hodnoty blízké nule a predikce by byla závislá pouze na hodnotě těchto parametrů, které by nebyly spolehlivé.

Po nahrání trénovací datové množiny do GeNIe jsme využili funkce `Learn New Network`, v nastavení jsme si zvolili Naive bayes s využitím všech proměnných a označení Událost EFS24 jako class variable. Nastavení je zobrazeno na obrázku 2.6.

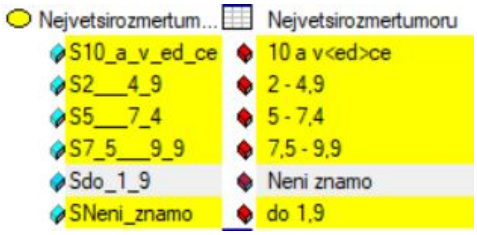
V dalším kroku je vygenerována struktura bayesovské sítě podle zmíněných předpokladů. Program GeNIe využívá na nalezení parametrů bayesovský přístup a EM algoritmus. Parametry se automaticky naučí z dat při provedení zmiňované funkce. Nicméně tyto parametry se dají znovu naučit zvolením jiných dat nebo jiných apriorních rozdělení pro parametry. Na výběr je buď rovnoměrné rozdělení, které nepřisuzuje žádnou důležitost dosud nalezeným parametrům a také žádnou specifickou znalost pravděpodobnostního rozdělení parametrů. Dále je možnost vybrat náhodnost, která vybere náhodně vygenerované čísla jako začátek algoritmu. Tato možnost se doporučuje hlavně pro data s latentními proměnnými. Třetí možnost je ponechat parametry originální, tato možnost se využívá pouze v případě, že jsou přidána nová data.

Na otestování výkonnosti modelu se využije funkce `Validate` v záložce Learning na testovací datovou množinu. Při tomto kroku, a i při učení nových para-



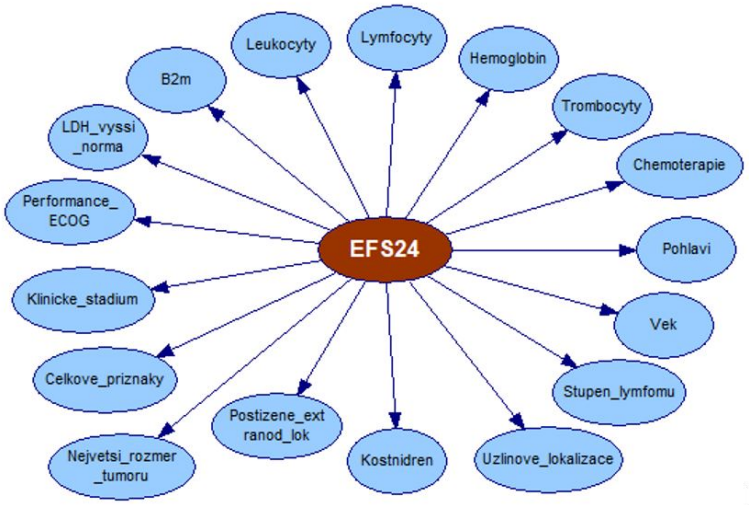
Obrázek 2.6: Nastavení pro získání Naivní bayesovské struktury

metrů, se program zeptá na shodu kategorií jednotlivých proměnných v síti a v datech. Při tomto příkladě je důležité vyměnit kategorii Není známo v proměnné Největší rozměr tumoru, která automaticky nabíhá špatně. Tato chyba je zobrazena na obrázku 2.7. Struktura naivní bayesovské sítě je zobrazena na obrázku 2.8.



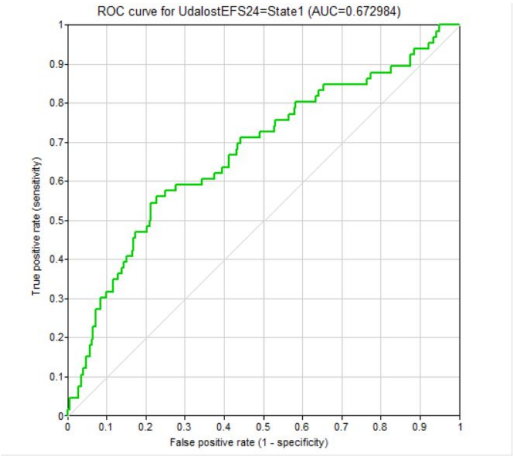
Obrázek 2.7: Výměna kategorií zvolené proměnné

Při využití funkce `Validate` pro testovou datovou množinu program spočítá jednotlivé pravděpodobnosti $P(Y = 1|X = E)$ a $P(Y = 0|X = E)$, tedy pravděpodobnost, že nastane, resp. nenastane, událost EFS24 na základě uvedených důkazů. Z těchto hodnot vykreslí ROC křivku, která je uvedena na obrázku 2.9. Plocha pod křivkou je 0,673, což je velmi vysoká hodnota, pokud vezmeme v



Obrázek 2.8: Struktura naivní bayesovské sítě

úvahu, že model je jednoduchý s přísnými předpoklady. Matice záměn společně se zvolenými statistikami je zobrazena v tabulce 2.7.



Obrázek 2.9: ROC křivka pro testovací datovou množinu u naivní bayesovské sítě

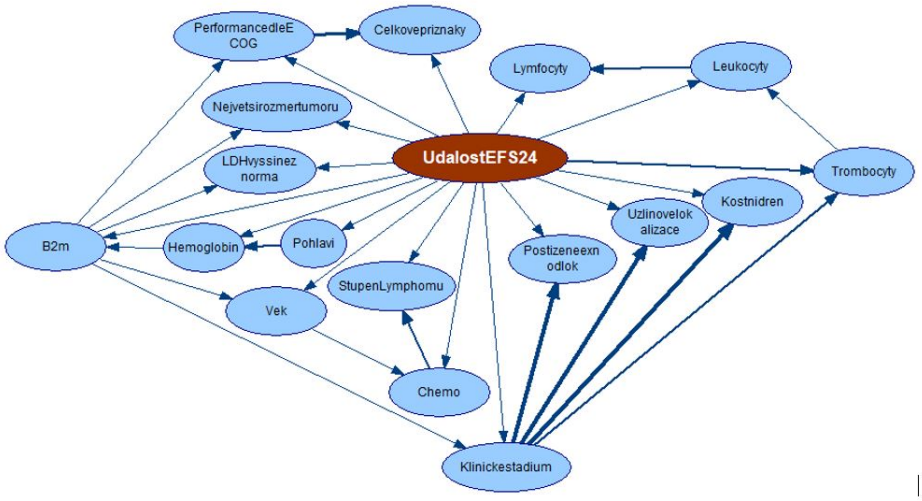
2.4.2. Stromově rozšířená bayesovská síť (TAN)

Stromově rozšířená bayesovská síť, tree augmented naive bayes network (TAN), nepředpokládá nezávislost vysvětlujících proměnných. Tyto proměnné bez naivní bayesovské struktury tvoří strom. Jedna proměnná je tedy kořen a ostatní

Skutečné hodnoty	Predikce		Statistika	
	0	1	Senzitivita	0.47
0	218	53	Specificita	0.80
1	35	31	F1	0.41

Tabulka 2.7: Matice záměr pro model naivní bayesovské sítě při zvoleném prahu 0,395

proměnné jsou vázány pravidlem, že mohou mít pouze dva rodiče. Algoritmus, který je založený na vytvoření struktury bayesovské sítě na základě hodnot podmíněně vzájemné informace proměnných je vysvětlen v kapitole 1.3.5. Obrázek 2.10 vykresluje jednu z možných TAN struktur na daných datech vytvořených v programu GeNIe. Tento typ struktury se jednoduše vybere v nabídce učení nových sítí. Je nutné si pamatovat, že při tak komplikovaných strukturách se vztahy tvoří někdy i opačným směrem, aby byly splněny podmínky k vytvoření acyklického směrového grafu. Hrany mezi proměnnými značí stochastické vztahy, ale ne nutně kauzální vztahy. V této struktuře je kořen stromu proměnná Pohlaví, protože má pouze jednoho rodiče, Událost EFS24, a strom se dále z ní rozvětvuje. Šířka hrany značí sílu statistické závislosti, která nemá vliv na výsledek při predikci, ale slouží jako další grafická informace ze sítě.



Obrázek 2.10: Struktura stromově rozšířené naivní bayesovské sítě

Výsledky modelu jsou o něco málo horší, kdy plocha pod křivkou nabývá hodnoty 0,653. Matice záměn a ostatní statistiky jsou zobrazeny v tabulce 2.8.

Skutečné	**Predikce**		**Statistika**	
hodnoty	0	1	Senzitivita	0.42
0	216	55	Specificita	0.80
1	38	28	F1	0.38

Tabulka 2.8: Matice záměr pro model stromově rozšířené naivní bayesovské sítě při zvoleném prahu 0,28

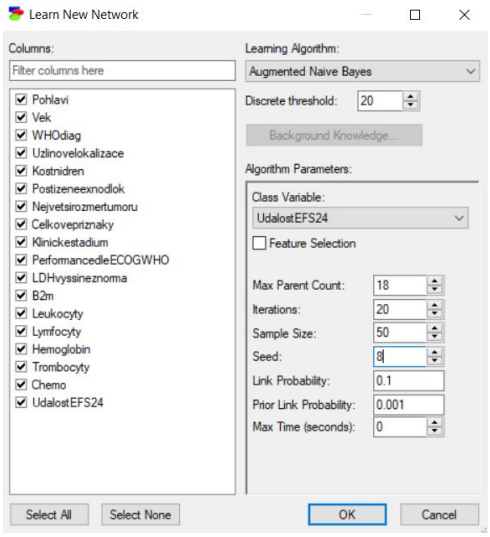
2.4.3. Rozšířená naivní bayesovská síť

Naivní bayesovská síť a TAN bayesovská síť mají omezení na nalezení vztahů mezi vysvětlujícími proměnnými hlavně z výpočetních důvodů. Pokud nejsou zadány žádné omezení na vztahy mezi vysvětlujícími proměnnými, a jsou vynuceny hrany z vysvětlované proměnné do všech ostatních, tak vzniká rozšířená naivní bayesovská struktura. Tento model může být výpočetně náročnější, ale může zachytit více vztahů, které se v tomto příkladě mohou vyskytovat.

Vytvoření této sítě v programu GeNIe probíhá následovně. Nejdříve se vynutí a zafixuje struktura naivní bayesovské sítě s určených klasifikačním uzlem a poté se tato síť ohodnotí podle vybrané skórovací funkce, MDL, a začne se s optimalizačním algoritmem na nalezení lepšího skóre pomocí operací: přidání, otočení či odebrání hrany. Optimalizační algoritmus využívá hill-climbing algoritmus s náhodnými restarty, které ale obsahují strukturu sítě s vynucenou naivní bayesovskou strukturou sítě, k vyhnutí se lokálnímu minimu. Do nastavení hledání struktury se zadává maximální počet iterací po kterých se algoritmus ukončí a výsledná síť odpovídá síti s nejlepším skóre. Podrobněji v kapitolách 1.3.3 a 1.3.5.

Na obrázku 2.11 je vidět nastavení této sítě. V nabídce naučit novou síť se vybere možnost rozšířené naivní bayesovské struktury, dále se může vybrat možnost *Výběr proměnných*, což ale pak dává vzniku síti, která je velmi málo propojená a tak je poté i málo výkonná v predikci. Problém predikce události EFS24 je velmi komplikovaný proces, které je dán mnoha faktory. Pokud tedy nezahrneme

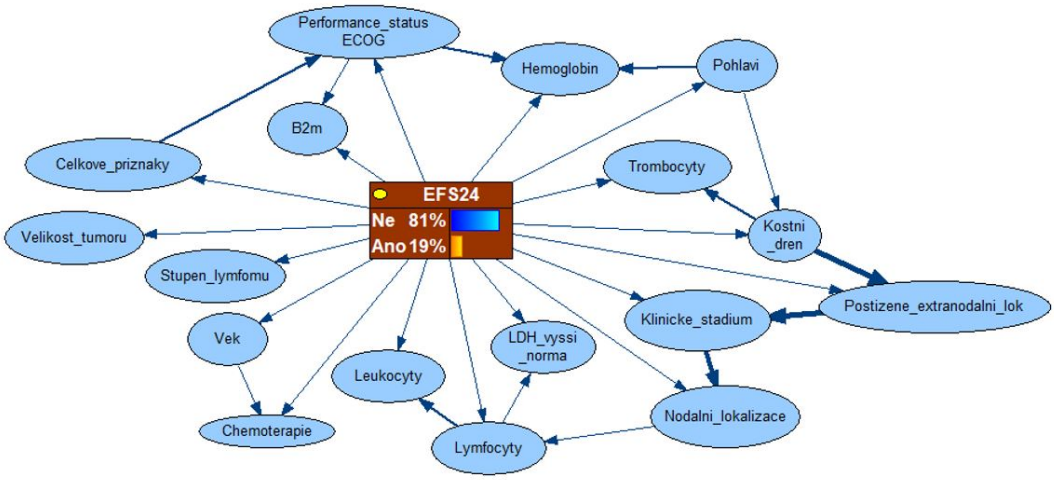
všechny dostupné, tak se ztrácí informace. Je možné stanovit maximální počet rodičů každého uzlu. Základní hodnota je 8, aby se předcházelo vyčerpání paměti. Parametr iterace udává počet nových startů algoritmu. Čím vyšší je množství iterací, tím se zvětšuje možnost nalezení lepší struktury, ale jelikož algoritmus hledá řešení v hyper-exponenciálním prostoru řešení, tak se s vyšším počtem zvyšuje i čas. Jako obvykle si zde člověk může vybrat mezi kratší dobou trvání nebo možnými lepšími výsledky. Dále se dá stanovit číslo na opětovné náhodné generování, apriorní pravděpodobnosti a také maximální čas.



Obrázek 2.11: Nastavení pro hledání rozšířené naivní bayesovské struktury sítě

Výsledky mohou být různé, protože problém je NP-obtížný, jak bylo zmíněné v kapitole 1.3.3. Jelikož v rámci algoritmu probíhají náhodné restarty, tak každý nový pokus může přinést o trochu jinou síť. V našem případě jako výsledná síť s nejvyšším skóre při různých možnostech nastavení vyšla síť na obrázku 2.12. Tato síť měla hodnotu AUC rovnu hodnotě 0,671, což je zobrazeno na obrázku 2.13. Matice záměn pro stanovený práh 0,295 (29,5% pravděpodobnost na výskyt události EFS24) a příslušné statistiky jsou zobrazeny v tabulce 2.9. Tento druh sítě se jeví jako nejlepší možné řešení pro predikci prognózy leukemických pacientů z daných dat pomocí tohoto druhu sítě. Nicméně, pro nalezení vhodné

struktury je potřeba stanovit správné apriorní pravděpodobnosti vztahů a také nechat algoritmus hledat dostatečně dlouho.



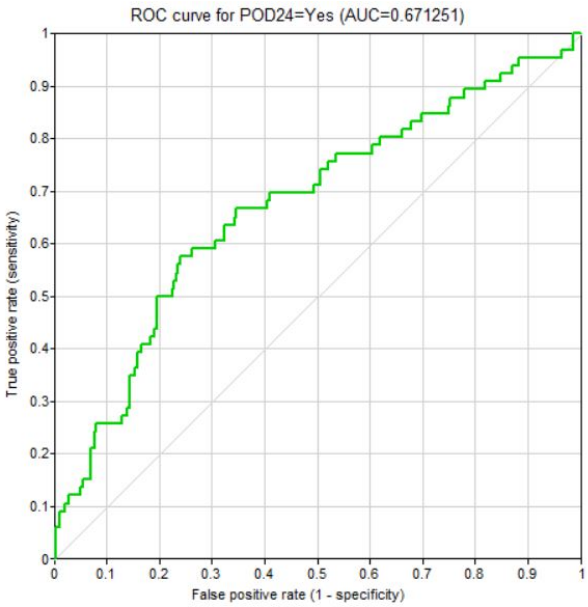
Obrázek 2.12: Výsledná struktura pro rozšířenou naivní bayesovskou síť

Skutečné hodnoty	**Predikce**		**Statistika**	
	0	1	Senzitivita	0.50
0	218	53	Specificita	0.80
1	33	33	F1	0.43

Tabulka 2.9: Matice záměr pro model rozšířené naivní bayesovské sítě při zvoleném prahu 0,295

2.4.4. Ostatní druhy učení struktury z dat

Pro některá data jsou klasické sítě, ty bez omezení nebo vynucení hran, lepší volbou pro pochopení systému. Nicméně, jak bylo řečeno v kapitole 1.3.5, tak tyto sítě poté nemusí být vhodné pro klasifikaci. Hlavně z důvodu, že Markov blanket vybrané proměnné EFS24 je vždy málo početná. Zde uvádíme některé příklady takovýchto sítí, které dané algoritmy nalezly při učení struktury z dat.



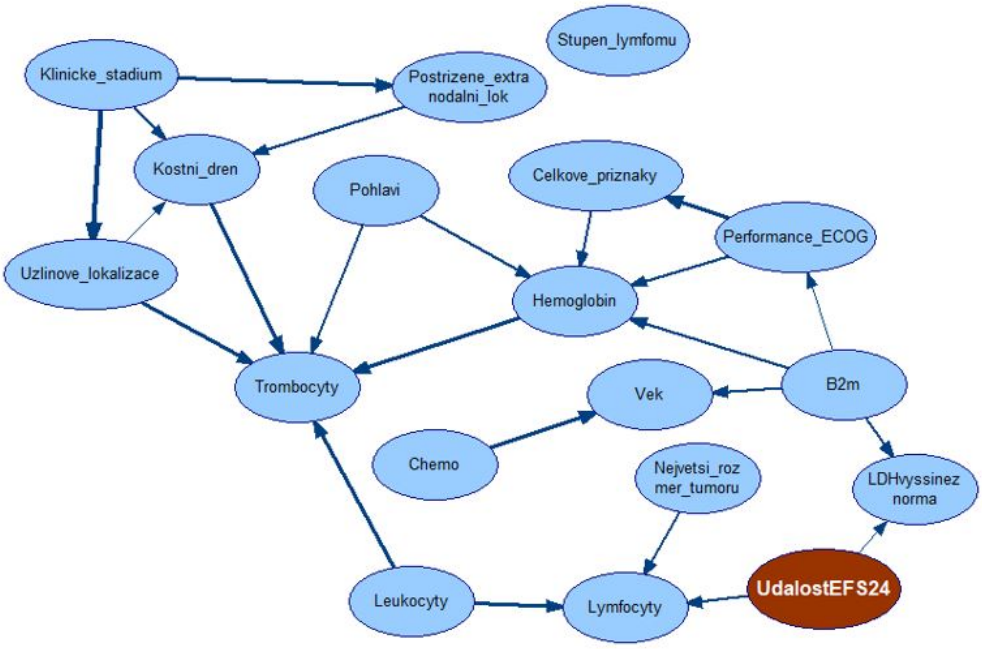
Obrázek 2.13: ROC křivka pro výslednou rozšířenou naivní bayesovskou síť

PC algoritmus

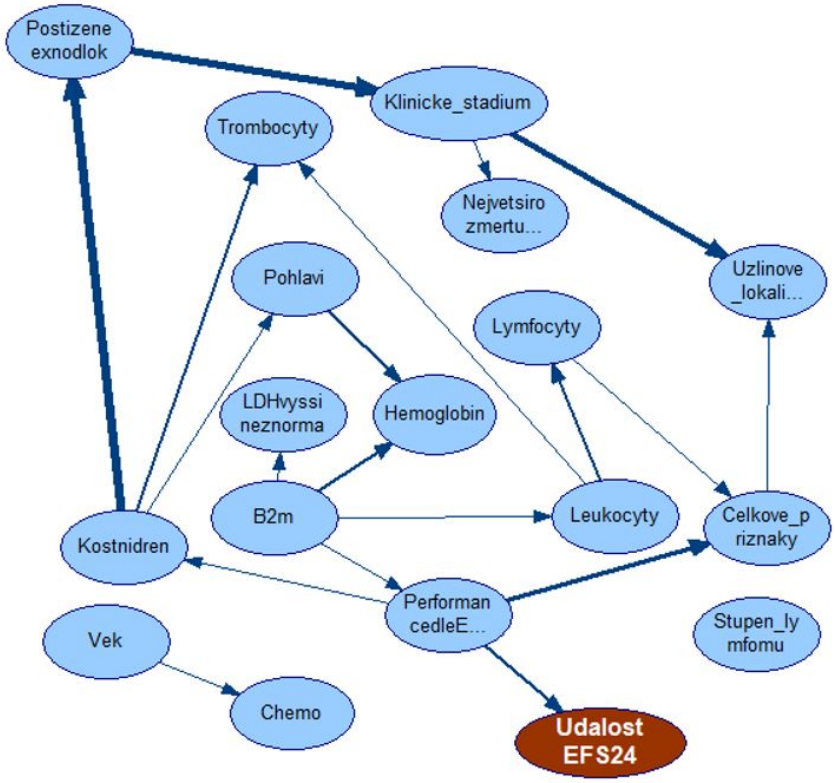
Tento algoritmus je založen na hledání vztahů podle klasických nezávislostních testů. Struktura je zobrazena na obrázku 2.14. Na predikci EFS24 je potřeba znát hodnoty proměnných: Leukocyty, Lymfocyty, Nevetší rozměr tumoru, B2m a LDH vyšší než norma. Nejde se proto divit, že AUC je pouhých 0,54, což je velmi blízko k náhodnému klasifikátoru.

Hill-climbing algoritmus bez omezení

Dále jsme vyzkoušeli, jak by vypadala síť s využitím hill climbing algoritmu s náhodnými restarty. Jak bylo řečeno, tak každé nové spuštění může najít jinou síť kvůli náročnosti problému. Jedna možná síť je uvedena na obrázku 2.15. Při vyhodnocení takové sítě je sice AUC 0,58, což je vyšší hodnota než v minulém případě, ale všechna pozorování jsou klasifikována jako 0, protože klasifikace závisí jen na parametrech CPD tabulky EFS24, která má pouze jednoho rodiče a to Performance dle ECOG. Tyto parametry pro stav 1 jsou při každé možnosti Performance dle ECOG nižší než pro stav 0. Taková síť je nevhodná pro klasifikace.



Obrázek 2.14: Bayesovská síť postavená na základě algoritmu PC



Obrázek 2.15: Bayesovská síť při využití hill climbing algoritmu bez omezení na strukturu

2.5. Porovnání využitých modelů

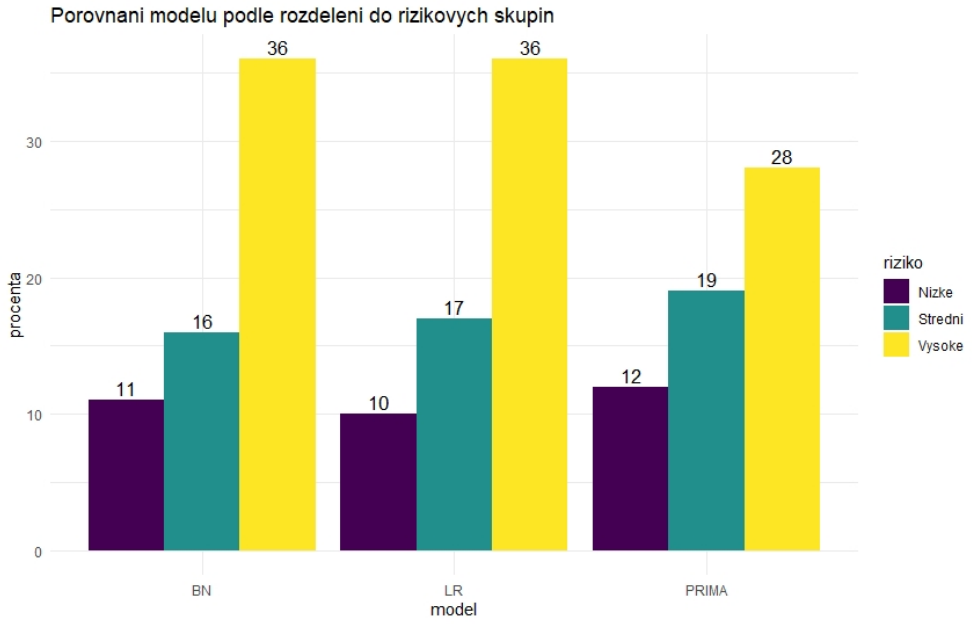
Vybrali jsme finální model bayesovské sítě, který odpovídá rozšířené naivní bayesovské síti, a srovnali s modelem logistické regrese a PRIMA-PI. Jelikož PRIMA-PI rozděluje pozorování do 3 skupin rizika a neprovádí binární klasifikaci, tak jsme museli využít k porovnání všech metod rozdělení pacientů do tří skupin i u logistické regrese a bayesovské sítě.

To bylo provedeno tak, že byly stanoveny hraniční hodnoty pro každou skupinu rizika pro predikční hodnotu tak, aby podíly každé skupiny rizika na celkovém počtu byly stejné jako u PRIMA-PI, což bylo 40 % z celkového počtu pozorování pro nízké riziko, 30 % z celkového počtu pozorování pro střední a 30 % z celkového počtu pozorování pro vysoké riziko. Seřazené predikované hodnoty jsme rozdělili do skupin tak, že nízké riziko obsahovalo data do 40. percentilu predikované hodnoty, střední riziko obsahovala data od 41. percentilu do 70. percentilu predikované hodnoty a vysoké riziko obsahovalo data od 71. percentilu se zbytkem hodnot. Tabulka 2.10 ukazuje výsledky a ty můžeme vidět graficky zobrazené na obrázku 2.16.

Lze vidět, že logistická regrese i bayesovská síť mají velmi podobné výsledky. Obě tyto metody ale mají lepší výsledky než PRIMA-PI, protože zastoupení pacientů u kterých nastala EFS24 je vyšší u nejrizikovější skupiny a nižší u nízkého rizika. Nicméně, tyto metody využívají více proměnných, a tak pracují s větším množstvím informací, proto dosahují lepších výsledků. Jak už to bývá, přesnost a jednoduchost jdou proti sobě.

Riziko	Model			Podíl kategorie
	PRIMA-PI	Logistická regrese	Bayesovská síť	
Nízké	12 %	10 %	11 %	40 %
Střední	19 %	17 %	16 %	30 %
Vysoké	28 %	36 %	36 %	30 %

Tabulka 2.10: Srovnávající tabulka všech tří metod - Procentuální výskyt EFS24=1 ve vybrané kategorii rizika pro srovnávané modely při zachování stejných podílu skupin na celkovém množství pacientů (poslední sloupec - podíl kategorie)



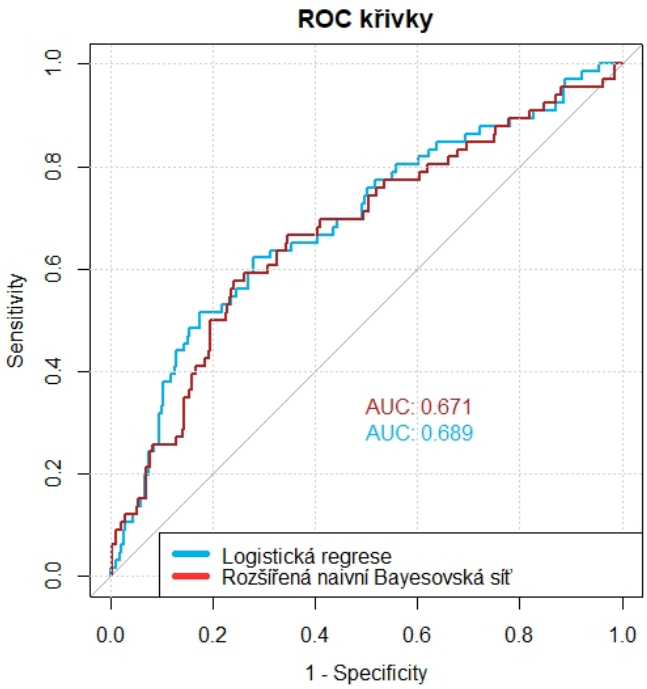
Obrázek 2.16: Grafické porovnání modelů pomocí rozdělení pacientů do rizikových skupin - Procentuální výskyt EFS24=1 ve vybrané kategorii rizika pro srovnávané modely při zachování stejných podílu skupin na celkovém množství pacientů

Při porovnání logistické regrese a bayesovské sítě pomocí ROC křivky je možné pozorovat podobnou výkonnost, což je zobrazeno na obrázku 2.17. To se projeví také u citlivosti a specifičnosti, při stanovení specifičnosti na 80 % má logistická regrese 51% citlivost a bayesovská síť má 50%. Jaké by tedy měly být benefity bayesovské sítě?

Hlavní předností bayesovské sítě by mělo být zobrazení vztahů, nicméně, pro rozšířenou naivní bayesovskou síť je struktura sítě zkreslena právě naivní bayesovskou strukturou, která vynucuje hrany z vysvětlované proměnné. Nicméně u většiny takto postavených sítí, se kterými jsme pracovali, se ustálily některé silné vztahy, které mají opodstatnění i z lékařského hlediska. Například: propojení Hemoglobinu a pohlaví, Leukocyty s Lymfocyty, Věk s Chemoterapií. Další velkou výhodou bayesovské sítě oproti logistické regresi je možnost vyhodnocení pravděpodobnosti EFS24 bez znalosti všech proměnných. U takto predikovaných hodnot neplatí daná ROC křivka, která počítá se znalostí hodnot všech

proměnných. K této výhodě se váže i další, která představuje jednoduché ovládání a grafické zpracování. Pro vyšetřujícího lékaře by stačilo ke každé proměnné kliknout na danou kategorii a sledovat, jak se mění výsledná pravděpodobnost. To by také mohlo pomoci s rozhodováním například o výběru typu chemoterapie, jelikož by se srovnával účinek různých chemoterapií na výslednou pravděpodobnost EFS24. Nicméně, tím, že model nemá velmi silnou přesnost, by tyto výsledky nemusely být vždy spolehlivé.

Žádná jiná metoda se ale neprojevila výrazně lépe, takže použití bayesovské sítě se jeví jako jedna z nadějných cest. Pro využití v reálném světě by bylo vhodné model natrénovat a otestovat na větších datech, díky čemuž by se zpřesnila hodnota parametrů a reálná výkonnost. Nicméně, pokud bychom chtěli vybrat model s maximální hodnotou AUC na testovacích datech, tak by jím byl model logistické regrese.



Obrázek 2.17: Srovnání ROC křivek pro model logistické regrese a rozšířené naivní bayesovské sítě pro testovací data

2.5.1. Porovnání modelů na konkrétním pacientovi

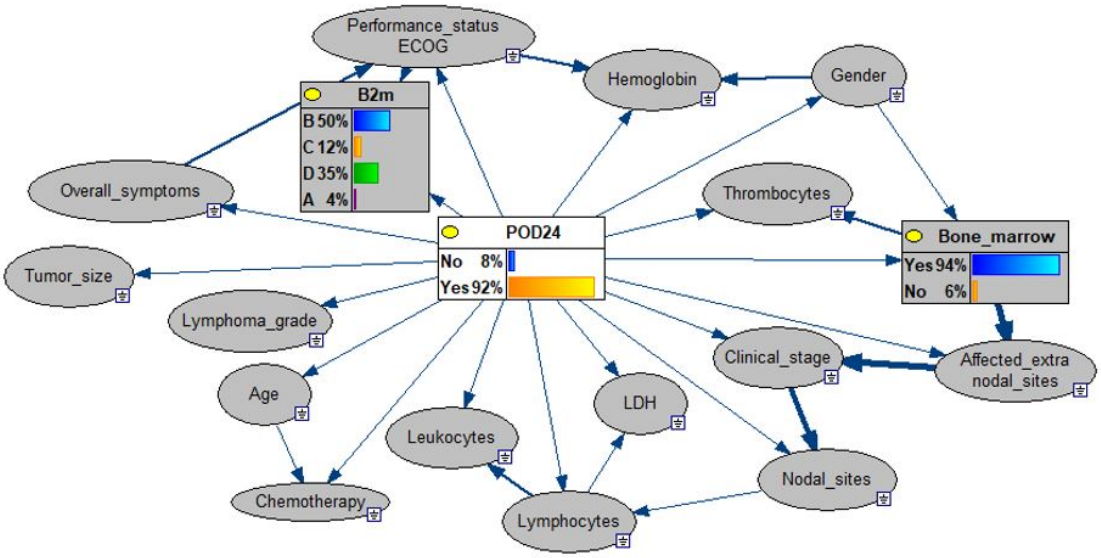
Abychom viděli přímý dopad modelů, tak se můžeme podívat na pacienta, který podle PRIMA-PI má nízké riziko EFS24, ale podle BN má vysoké riziko EFS24. Hodnoty proměnných vstupující do modelu jsou zobrazeny v tabulce 2.11. Pacient prodělal EFS24 a bayesovská síť mu přidělila pravděpodobnost události EFS24 rovnu 91 %. Naproti tomu, jeho hodnota B2m byla nízká a kostní dřeň nebyla infikována, proto pacient spadal do nízké kategorie rizika podle PRIMA-PI.

Proměnná	Pacient
Pohlaví	Žena
Věk	63
Stupeň lymfomu	FL 3A
Nodální lokalizace	4
Kostní dřeň	Ne
Postižené extranodální lokalizace	1
Velikost tumoru	Není známo
Celkové příznaky	Ano
Klinické stadium	IV
Performance dle ECOG	3-4
LDH vyšší než norma	Ano
B2m	2.4
Leukocyty	13.46
Lymfocyty	0.67
Hemoglobin	111
Trombocyty	160
Chemoterapie	Chop
P(POD24=1)	91 %
PRIMA-PI	Nízké riziko
POD24	Ano

Tabulka 2.11: Popis sledovaného pacienta

Dále se můžeme podívat, jak se výsledek změní, pokud síti neposkytneme hodnoty proměnné B2m a Kostní dřeň. Výsledek je pravděpodobnost události EFS24 rovna 92 %. Tento výstup je zobrazen na obrázku 2.18. Tento postup by nebyl tak jednoduchý při logistické regresi, jelikož tam bychom museli nějaké

hodnoty těchto proměnných zadat. Možné řešení by byla vhodná imputace hodnot na základě ostatních hodnot proměnných a nebo jednoduché řešení imputace mediánem nebo průměrem. Při analyzování výsledků modelu bayesovské sítě bez využití proměnných B2m a Kostní dřeň můžeme sledovat pokles výkonnosti, ale tento pokles není výrazný. AUC hodnota klesne o 2 procentní body na hodnotu 0,65 a senzitivita o 3 procentní body na 0,47 při 80% specificitě.



Obrázek 2.18: Výsledná síť (EN) pro sledovaného pacienta bez poskytnutí informace o hodnotě proměnné B2m a Kostní dřeň (POD24 odpovídá EFS24)

2.5.2. Ilustrace využití bayesovské sítě při volbě protokolu léčby

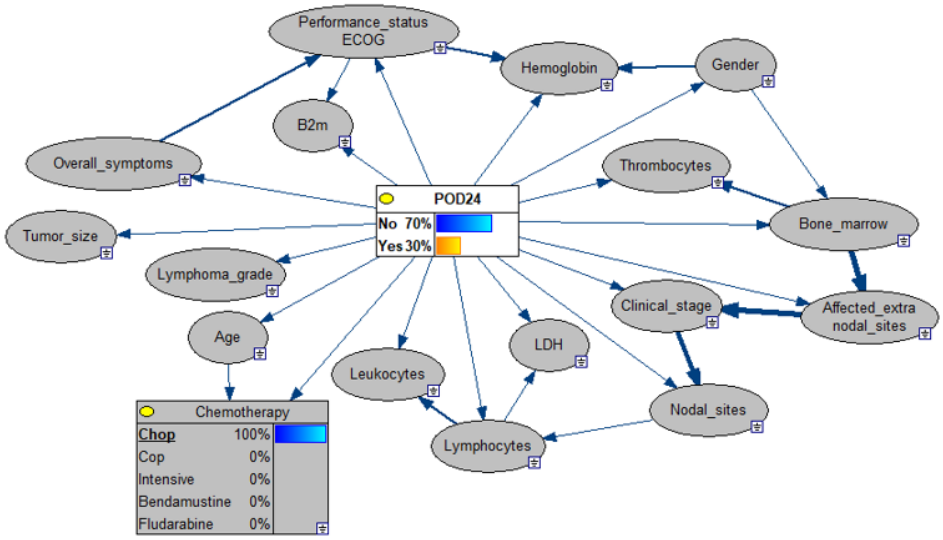
U jiného pacienta se můžeme podívat na změnu, která nastane při změně druhu chemoterapie. Popis parametrů pacienta je ukázán v tabulce 2.12. U tohoto pacienta se můžeme podívat na rozdíl při využití léčby Chop, kterou dostal, a také jaké by byly výsledky predikce prognózy při léčbě Cop. Tento rozdíl na zobrazen na obrázku 2.19 a 2.20.

Tato aplikace by mohla mít pro lékaře velmi dobré praktické využití v reálném životě, aby jim pomohla jako statistický podklad při výběru typu léčby. Nicméně,

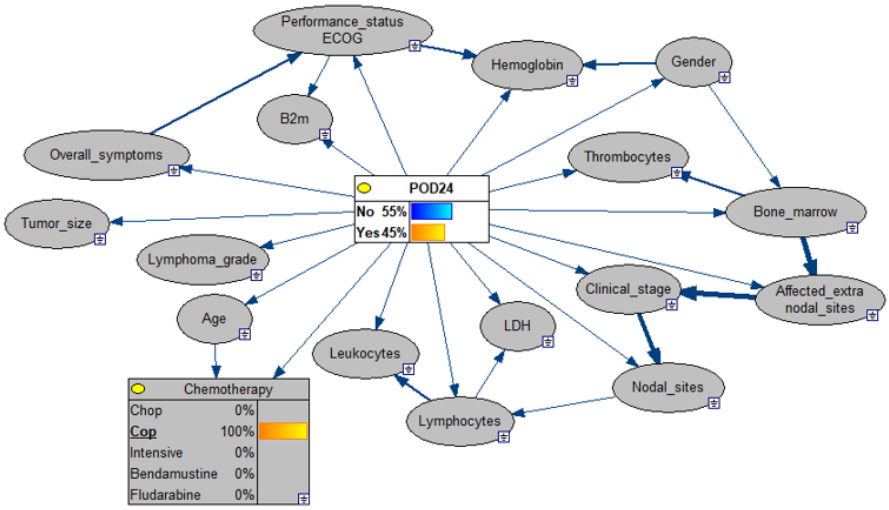
většina pozorování z vybrané datové množiny měla typ léčby Chop, přesněji 1039 z 1348, viz kapitola 2.1.3. To způsobilo, že síť neměla dost dat ze kterých by se mohla naučit dostatečně přesné parametry, které by dávaly přesné výsledky při různých typech léčby. Nicméně, lze na tomto příkladě ilustrovat možné využití bayesovské sítě například u jiných problémů nebo u situací s větším množstvím dat.

Proměnná	Pacient
Pohlaví	Muž
Věk	48
Stupeň lymfomu	FL 2
Nodální lokalizace	9
Kostní dřeň	Ano
Postižené extranodální lokalizace	1
Velikost tumoru	5.5-7.4
Celkové příznaky	Ne
Klinické stadium	IV
Performance dle ECOG	0
LDH vyšší než norma	Ano
B2m	5.56
Leukocyty	4.7
Lymfocyty	0.82
Hemoglobin	131
Trombocyty	103
Chemoterapie	Chop
P(POD24=1)	30 %
PRIMA-PI	Vysoké
POD24	Ano

Tabulka 2.12: Popis sledovaného pacienta pro rozdíl výsledků léčby



Obrázek 2.19: Bayesovská síť (EN) pro popisovaného pacienta při využití léčby Chop a jeho výsledek predikce prognózy (POD24 odpovídá EFS24)



Obrázek 2.20: Bayesovská síť (EN) pro popisovaného pacienta při využití léčby Cop a jeho výsledek predikce prognózy (POD24 odpovídá EFS24)

Závěr

Cílem této práce bylo představit teorii bayesovských sítí a aplikovat ji na vytvoření modelu pro predikci prognózy leukemických pacientů v ČR na základě dat poskytnutých Kooperativní lymfomovou skupinou ČR. Až 20 % pacientů s folikulárním lymfomem postihuje progrese, relapse nebo smrt do 2 let. Proto bylo cílem navrhnout model, který by co nejpřesněji udával pravděpodobnost takovéto události už při počáteční léčbě ke stanovení správného protokolu léčby.

Práce seznamuje čtenáře se základní teorií bayesovských sítí, ale také s teorií, jak využít bayesovské sítě při klasifikační úloze. Bayesovské sítě se vytvářejí k docílení 3 základních cílů: odhad pravděpodobnostního rozdělení, stanovení pravděpodobnostních vztahů nebo ke klasifikace vybrané proměnné. Tyto cíle někdy nemusejí jít spolu dohromady, což bylo ověřeno u našich dat, kdy některé struktury sítě by možná lépe reflektovaly skutečný proces systému, ale nebyly vhodné pro klasifikaci vybrané proměnné.

Byly proto vyzkoušeny různé možnosti bayesovských sítí a byl vybrán nejvhodnější model ke klasifikaci nových pozorování. Modely byly budovány na trénovacích datech a jejich výkonnost byla ověřována na testovacích datech. Výsledný model byl srovnán s ostatními modely využívanými v této oblasti, tedy logistickou regresí a PRIMA prognostickým indexem. Logistická regrese a bayesovská síť měly lepší výsledky než PRIMA-PI, což bylo očekáváno, protože PRIMA-PI využívá informaci pouze ze dvou proměnných. Při porovnání modelů při rozdělení pozorování do 3 skupin rizika, měly logistická regrese a bayesovská síť o 8 procentních bodů vyšší riziko v nejrizikovější kategorii než PRIMA-PI. Dále při porovnání logistické regrese a bayesovské sítě pomocí AUC, citlivosti

a specifičnosti, měla logistická regrese jen o 1,8 procentního bodu lepší AUC hodnotu a o 1 procentní bod lepší citlivost při 80% specificitě.

Hlavní předností bayesovské sítě je její srozumitelná grafická reprezentace a vyhodnocení pravděpodobnosti události bez znalosti všech proměnných. U takto predikovaných hodnot ale neplatí daná ROC křivka, která počítá se znalostí hodnot všech proměnných. Z těchto výhod plyne možnost sledovat změnu pravděpodobnosti při změně hodnot jednoduchým a srozumitelným způsobem, například u vlivu různých druhů léčby.

Úloha klasifikace EFS24 se jeví jako komplikovaný problém, u kterého se i přes použití různých modelů nedosahuje velmi kvalitní výkonnosti. Je možné, že tento problém je ovlivněn jinými faktory, které nebyly zkoumány. Přesto využití bayesovských sítí se jeví jako vhodná alternativa, jelikož žádná jiná zkoumaná metoda se neprojevila výrazně lépe.

Obsah této práce, zejména praktická část, tedy zpracování dat do podoby bayesovské sítě a porovnání s logistickou regresí a PRIMA-PI, byl zadán a později vybrán k ústní prezentaci do 62. ročníku kongresu Americké hematologické asociace (62nd American Society of Hematology annual meeting), která je s 30.000 účastníky a více než 4.000 sděleními největší a nejprestižnější hematologickou událostí na světě. Abstrakt byl již publikován v online suplementu časopisu Blood. Práce také zaujala japonské kolegy natolik, že se rozhodli tuto práci přeložit do japonštiny pro své publikum. Tento článek je přiložen v Příloze 2.

Příloha 1

Součástí této práce je výsledná rozšířená naivní bayesovská síť v elektronické formě z programu GeNIe s názvem „ANBN.xdsl". Tento soubor obsahuje strukturu sítě a výsledné parametry a je uložen v přílohách této diplomové práce.

Příloha 2

Další přílohou této práce je obsah abstraktu, přeložený do japonštiny, který byl prezentován na 62. ročníku kongresu Americké hematologické asociace, obrázek 2.21.

#375

Bayesian Network Modelling as a New Tool in Predicting of the Early Progression of Disease in Follicular Lymphoma Patients

濾胞性リンパ腫(FL)患者における病勢の早期進行を予測する新たなツールとしてのベイジアンネットワークモデル

Vít K. Procházka, et al. University Hospital Olomouc, Czech Republic

背景・目的

高腫瘍量の初期FL患者では約20%が治療開始から24カ月以内に再発(POD24)するとされている。POD24の予測指標として、現在、ロジスティック回帰分析(LR)によるFLIPIやPRIMA-PIなどの予後予測スコアが使用され、その特異度は47～86%とされている。一方、ベイジアンネットワーク(BN)はLRと比べて、変数間の複雑な因果関係を把握でき、過剰適合の回避や欠損データの補完も可能といった利点がある。そこで今回、未治療のFL診断時のデータを用いて、POD24のリスク予測モデルを構築し、LRとBNのPOD24予測精度を比較した。加えて、既存のPRIMA-PIコホートも予測精度比較のため活用した。

方法

2000年～2016年にチェコ国内で初発FLと診断され、リツキシマブ(R)を含む治療レジメンで導入療法を施行し、チェコリンパ腫研究グループ(CLSG)レジストリに登録された1,394例を対象とした。POD24は導入療法開始から24カ月以内の再発、進行及び治療法の変更と定義した。

結果

対象1,394例の患者背景は、全体の44.5%が60歳を上回り、PRIMA-PI群に比べ約10%多かった。StageⅢ-Ⅳは85.9%で、導入療法はR-CHOP療法が76.8%と最も多く、R維持療法は67.1%で施行されていた(表)。観察期間中央値7.64年において、484例(34.7%)が進行又は再発し、316例(22.6%)が死亡した。POD24は266例(19.0%)に認められ、診断後5年の無増悪生存率は64.2%、全生存率は86.4%であった。LRモデルの構築に際しては過剰適合を回避するため、データを訓練データセット(75%)とテストセット(25%)に分類した。BN(Augmented Naïve Bayes分類器)に関してはPOD24を予測する全因子とリンクを張る一方で、年齢と性別へのリンクは禁止し、その他のネットワークの構成はデータから推測した。LRとBNの両モデルにおいてROC曲線下面積(AUC)よりPOD24に対するcut off値を29.5%と設定し、感度、特異度、陰性的中率(NPV)、陽性的中率(PPV)を評価した。BNモデルのAUCは0.67で、29.5%をcut off値とした場合の感度は50%、特異度は80%、NPV、PPVはそれぞれ87%、38%と、LRモデルと遜色のない結果が得られ、HighリスクPRIMA-PIよりも予測能が高かった(図)。

結論

POD24の予測ツールとしてBNモデルは最適とされるLRに劣らないツールである。BNは変数間の複雑な因果関係を可視化し、個別に予後を予測することができる。また、BNは欠損データがあってもリスクの算出ができ、治療効果の予測を可能にすると考えられる。

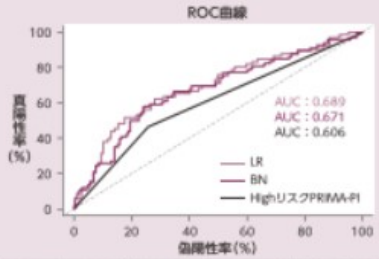
表 患者背景

		CLSG(n=1,394)	PRIMA-PI*(n=1,135)
年齢>60歳		44.5%	35%
男性		40.8%	52%
ECOG PS>1		9.9%	4%
β_2ミクログロブリン>3mg/L		33.1%	30%
臨床病期StageⅢ-Ⅳ		85.9%	90%
骨髄浸潤		48.9%	56%
導入療法	R-CHOP	76.8%	74%
	R-CVP	12.4%	22%
	R-フルダラビン	3%	4%
	R-ベンダムスチン	4.7%	0%
	強化型化学療法	3.3%	0%
R維持療法		67.1%	50%

*Bachy E, Seymour JF, Feugier P, et al. J Clin Oncol. 2019;37(31):2815-2824.

Presented at ASH - December 5-8, 2020 (virtual)

図 LRモデルとBNモデルの予測精度



予測モデル	AUC	感度	特異度	陰性的中率	陽性的中率
H-PRIMA-PI*	0.61	0.47	0.74	0.85	0.31
LR(ad hoc)	0.69	0.51	0.82	0.87	0.41
BN	0.67	0.50	0.80	0.87	0.38

*HighリスクPRIMA-PI

Presented at ASH - December 5-8, 2020 (virtual)

Obrázek 2.21: Přeložený abstrakt z ASH kongresu japonskými kolegy

Literatura

[1] Procházka, V., et al.: *Folikulární lymfom* Mladá fronta, a.s., 2017.

[2] Bart Baesens, Geert Verstraeten, Dirk Van den Poel, Michael Egmont-Petersen, Patrick Van Kenhove, Jan Vanthienen: *Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers* European Journal of Operational Research, Volume 156, Issue 2, Pages 508-523, 2004. Dostupné z: https://doi.org/10.1016/S0377-2217(03)00043-2

[3] Youtian Du, Feng Chen, Wenli Xu and Yongbin Li: *Recognizing Interaction Activities using Dynamic Bayesian Network* 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, pp. 618-621, 2006.

[4] H. Suk, B. Sin and S. Lee: Recognizing hand gestures using dynamic Bayesian network 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, Amsterdam, 2008.

[5] Ji Z., Xia Q., Meng G.: *A Review of Parameter Learning Methods in Bayesian Network.* In: Huang DS., Han K. (eds) Advanced Intelligent Computing Theories and Applications. ICIC 2015. Lecture Notes in Computer Science, vol 9227. Springer, Cham.

[6] Gámez, J.A., Mateo, J.L. & Puerta, J.M.: *Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood.* Data Min Knowl Disc 22, 106–148, 2011. Dostupné z: https://doi.org/10.1007/s10618-010-0178-6

[7] Česká onkologická společnost České lékařské společnosti J. E. Purkyně, 2021 [online]. LINKOS [Cit. 28.3.2021]. Dostupné z: https://www.linkos.cz/pacient-a-rodina/onkologicke-diagnozy/lymfomy-c81-85/

[8] Kooperativní lymfomová skupina, 2019 [online]. KLS [Cit. 28.3.2021]. Dostupné z: https://www.lymphoma.cz/

[9] Tom Fawcett: *An introduction to ROC analysis.* Pattern Recognition Letters, Volume 27, Issue 8, Pages 861-874, 2006. Dostupné z: http://https://doi.org/10.1016/j.patrec.2005.10.010

[10] Paul D. Allison: *Logistic Regression Using SAS: Theory and Application* SAS Institute, 2012.

[11] Bachy E, Maurer MJ, Habermann TM, et al.: *A simplified scoring system in de novo follicular lymphoma treated initially with immunochemotherapy* Blood. 2018; 132(1):49-58. Dostupné z: doi:10.1182/blood-2017-11-816405

[12] Daphne Koller, Nir Friedman: *Probabilistic Graphical Models: Principles and Technique* The MIT press, London, 2009.

[13] Gebhardt J., Kruse R.Gebhardt J., Kruse R.: *Background and perspectives of possibilistic graphical models*. In: Gabbay D.M., Kruse R., Nonnengart A., Ohlbach H.J. (eds) Qualitative and Quantitative Practical Reasoning. FAPR 1997, ECSQARU 1997. Lecture Notes in Computer Science, vol 1244. Springer, Berlin, Heidelberg, 1997. Dostupné z: https://doi.org/10.1007/BFb0035616

[14] Dupač, V., Hušková, M.: *Pravděpodobnost a matematická statistika* Karolinum, Praha, 1999.

[15] Scutari, M.: *Bayesian Network Constraint-Based Structure Learning Algorithms: Parallel and Optimised Implementations in the bnlearn R Package* Journal of Statistical Software (2017), 77(2), 1-20. Dostupné z: https://arxiv.org/pdf/1406.7648.pdf

[16] Gupta A, Slater JJ, Boyne D, et al.: *Probabilistic Graphical Modeling for Estimating Risk of Coronary Artery Disease: Applications of a Flexible Machine-Learning Method*. Medical Decision Making. 2019;39(8):1032-1044. Dostupné z: https://doi.org/10.1177/0272989X19879095

[17] Scutari, M., Vitolo, C. & Tucker: *A. Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation*. Stat Comput 29, 1095–1108 (2019). Dostupné z: https://doi.org/10.1007/s11222-019-09857-1

[18] RAYMOND, Yeung W. Information Theory and Network Coding. 1. vyd. New York: Springer, 2008. 600 s. ISBN 0-387-79233-3.

[19] Yun Zhou, Norman Fenton, Martin Neil: *Bayesian network approach to multinomial parameter learning using data and expert judgments* International Journal of Approximate Reasoning, Volume 55, Issue 5, 2014, 1252-1268. Dostupné z: https://doi.org/10.1016/j.ijar.2014.02.008.

[20] Friedman, N., Geiger, D. & Goldszmidt, M.: *Bayesian Network Classifiers* Machine Learning 29, 131–163 (1997). Dostupné z: https://doi.org/10.1023/A:1007465528199

[21] Chow, C. K. &C. N. Liu: *Approximating discrete probability distributions with dependece trees* IEEE Trans. on Information Theory, 14, 462-467, 1968.

[22] Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. Dostupné z: https://CRAN.R-project.org/package=stargazer.

[23] BayesFusion, LLC: Modely využité v této práci byly zpracovány softwarem GeNie Modeler, který je poskytován zdarma pro využití na akademický výzkum a vzdělání od BayesFusion, LLC. Dostupné z: https://bayesfusion.com/