Onemocníme cukrovkou? Co vše může k rozvoji tohoto onemocnění přispívat? (Metoda hlavních komponent)

V této práci se budu onemocněním cukrovkou, tedy diabetus mellitus, a faktory, které se spolupodílejí na rozvoji tohoto závažného onemocnění. Diabetes představuje velmi vážný problém, neboť počet osob s diabetem neustále narůstá. Nejenže výrazným způsobem ovlivňuje životy milionů lidí, ale dokonce zvyšuje celkovou úmrtnost a zkracuje délku dožití lidí s tímto onemocněním (Olorunfemi et al., 2025).

Počet lidí trpících diabetem (II. typu) tedy neustále roste. Podle údajů poskytnutých Ústavem zdravotnických informací a statistiky z roku 2021 žije v České republice téměř 1 100 000 osob s diabetem. K rozvoji onemocnění přispívají vnější i vnitřní faktory, v potaz je nutno brát rovněž genetickou zátěž. Mezi významné rizikové faktory pro vznik tohoto onemocnění patří nadváha a obezita (zejména kumulace tuku v oblasti břicha), nedostatek pohybu a nezdravá strava. Redukcí hmotnosti u jedinců s nadváhou či dokonce obezitou a dodržováním zásad zdravé životosprávy je možné riziko onemocnění značně eliminovat (IKEM, 2021).

Pro účely studijního předmětu Vícerozměrné statistické metody jsem vybrala dataset v databázi Kaggle[1][2] a zvolila **metodu hlavních komponent** (Principal Components Analysis, PCA). Dataset byl pro naše výpočty redukován. Byli vybráni pouze respondenti, kteří onemocněním cukrovkou trpí a byly odstraněny některé přebytečné sloupce (zda si respondent platí zdravotní pojištění, výše příjmu apod.). Výzkumný soubor je tvořen 35 346 respondenty. Dále zde máme sloupce BP, tedy krevní tlak (vysoký reprezentován 1, normální a nižší 0), vysoký cholesterol HighChol (ano 1, ne 0), BMI, kontrola cholesterolu v uplynulých 5 letech Cholcheck (reverzní položka, kuřák či nekuřák (Smoker 1, nonsmoker 0), mrtvice Stroke, infarkt HeartDiseaseorAttack, fyzická aktivita PhysActivity (reverzní položka), ovoce a zelenina Fruits, Veggies (opět reverzní položky), vysoký příjem alkoholu HvyAlcoholComp, celkové zdraví GenHlth, potíže s chůzí do schodů DiffWalk, pohlaví Sex, věk Age. Dataset obsahuje binární proměnné (např. HighBP, HighChol) a numerické proměnné (např. BMI, Age), PCA lze aplikovat i na tento typ dat. Byla použita rotace

[1] Teboul, A. (n.d.). *Diabetes health indicators dataset*. Kaggle. Dostupné z https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.

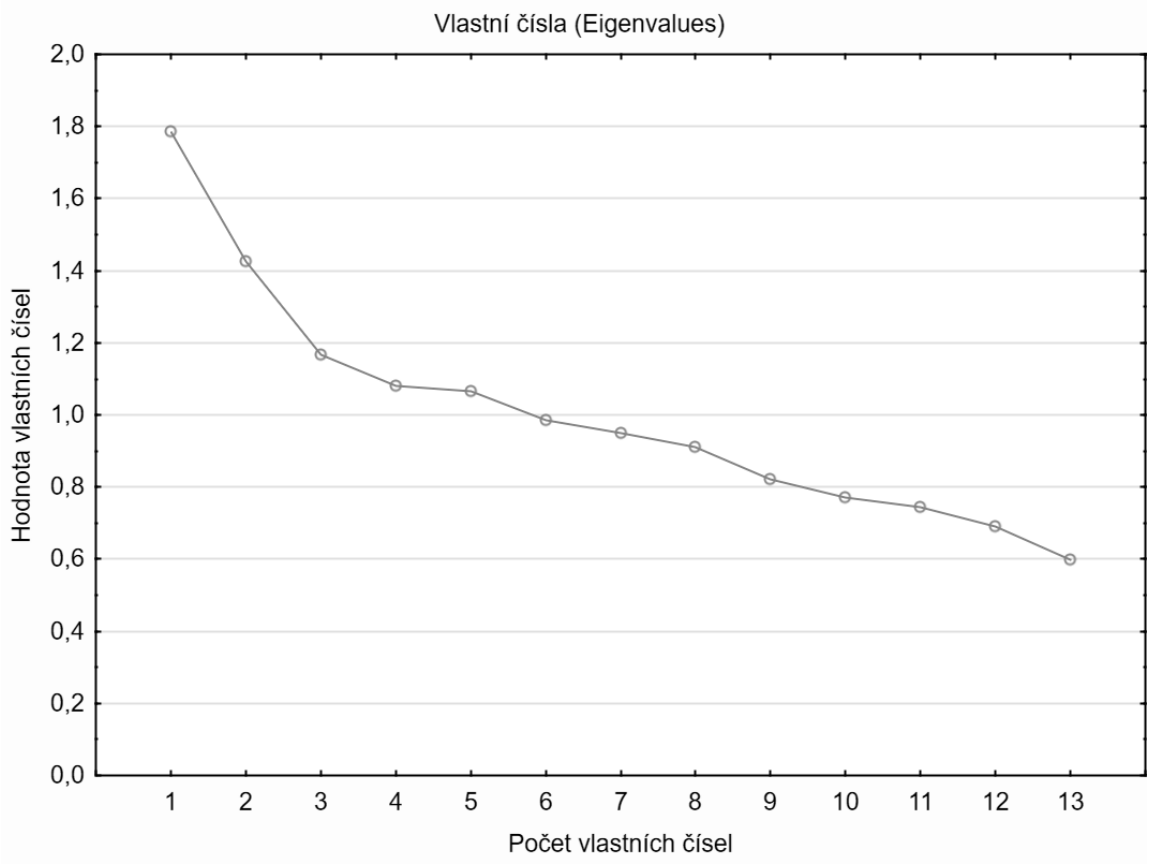
[2] Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4?i=465.

Varimax normalized. Varimax je jedna z nejběžněji používaných rotací usnadňující interpretaci faktorů nebo komponent v analýze faktorů a metodě hlavních komponent.

Nejprve bylo třeba otočit **reverzní položky.** To provedeme pomocí vzorce R = max(x) + min(x) – x. Mezi reverzní položky patřila konzumace ovoce a zeleniny (alespoň 1 ks denně), dále pravidelné provozování fyzické aktivity (v předešlých 30 dnech), kontrola cholesterolu (alespoň jednou za posledních 5 let). Celkové zdraví nepředstavovalo reverzní položku, neboť 1 reprez. velmi dobré zdraví a 5 velmi špatné.

Dále jsem si zobrazila **sutinový graf**, aby bylo možné určit počet hlavních komponent. Vlastní čísla s hodnotou vyšší než 1 jsou v počtu 5, dle grafu by bylo možné určit počet hlavních komponent také jako 3 (poté se graf láme a dále pozorujeme tzv. „suť“).

3 hlavní komponenty by vysvětlily přibližně 33,69 % celkového rozptylu, zatímco 5 hlavních komponent vysvětlí 50,21 % celkového rozptylu. Z těchto hodnot je patrné, že struktura modelu je poněkud složitější a komponenty vysvětlují pouze asi polovinu rozptylu.



Poté spočítám faktorové náboje jednotlivých položek a komunalitu. V tabulce č. 1 jsou znázorněny faktorové náboje a rovněž hodnoty komunality těchto položek.

Tabulka č. 1: Faktorové náboje a komunalita

Proměnná	Faktorové náboje					
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Komunalita
HighBP	0,17	0,08	0,00	0,73	0,02	0,56
HighChol	0,05	0,01	0,05	0,68	-0,20	0,51
BMI	0,03	0,83	-0,02	0,13	0,07	0,70
Smoker	0,26	-0,07	0,05	-0,03	-0,65	0,49
Stroke	0,58	-0,09	-0,01	-0,06	-0,02	0,35
HeartDiseaseorAttack	0,62	-0,14	-0,04	0,12	-0,13	0,44
HvyAlcoholConsump	-0,15	0,04	-0,02	0,02	-0,72	0,54
DiffWalk	0,63	0,40	0,04	0,07	0,03	0,57
Age	0,40	-0,56	-0,05	0,25	0,17	0,56
R-PhysActivity	0,41	0,34	0,33	-0,03	0,07	0,40
R-Fruits	-0,07	0,08	0,74	0,01	-0,16	0,58
R-Veggies	0,05	-0,06	0,79	0,02	0,11	0,64
R-CholCheck	0,12	0,05	0,02	-0,36	-0,18	0,18
Expl.Var	1,59	1,32	1,29	1,23	1,10	
Prp.Totl	0,12	0,10	0,10	0,09	0,08	

Faktor 1 je nejvíce sycen položkami Stroke (mrtvice), DiffWalk (potíže při chůzi do schodů) a R-PhysActivity (reverzní pol. fyzická aktivita). Souhrnně tedy Faktor 1 vypovídá o tělesné kondici. Položka fyzické aktivity R-PhysActivity však zároveň významně sytí i druhý a třetí faktor. DiffWalk rovněž do značné míry sytí i Faktor 2. Faktor 2 je výrazně sycen položkou BMI, kde je hodnota dokonce 0,83. Souhrnně souvisí Faktor 2 s tělesnou hmotností a rovněž kondicí. Faktor 3 je sycen především položkami R-Fruits a R-Veggies, tedy reverzními pol. konzumace ovoce a zeleniny. Lze jej popsat jako zdravá strava. Faktor 4 je výrazně sycen HighBP a HighChol, tedy vysoký tlak a cholesterol. Tyto dva indikátory představují zdraví kardiovaskulárního systému a bývají obvykle udávány společně. Faktor 5 sytí položky Smoker (kouření) a HvyAlcoholConsump (konzumace alkoholu). Tyto dvě závislosti bývají rovněž obvykle zmiňovány společně, můžeme je označit souhrnně jako legální drogy nebo závislost na alkoholu nebo nikotinu. Položka kontrola cholesterolu CholCheck nesytí významně žádný z hlavních pěti faktorů a její komunalita je též velmi nízká (0,18). Komunalita vyjadřuje podíl, jakým faktory přispívají do celkového rozptylu, tedy podíl variability každé proměnné, který je vysvětlen společnými faktory. Komunalita zjednodušeně řečeno říká, jak dobře je daná proměnná "vysvětlena" faktory v modelu faktorové analýzy. Komunalita blízká 0 znamená, že faktory vysvětlují velmi málo variability dané proměnné, tedy že většina variability je specifická pro tuto proměnnou

a není sdílena s ostatními. Položku CholCheck by bylo proto vhodné vyřadit, protože nesouvisí tolik s ostatními položkami.

Závěr

Za použití metody hlavních komponent bylo zjištěno, že vznik onemocnění diabetem sytí 5 komponent. Těmi jsou tělesná kondice, hmotnost, zdravá strava, zdraví kardiovaskulárního systému a nepřítomnost závislosti. Položky fyzická aktivita PhysActivity a potíže s chůzí do schodů DiffWalk sytily zároveň faktor 1 i faktor 2. Položku CholCheck, tedy kontrola cholesterolu v předešlých 5 letech, nesouvisí tolik s ostatními položkami, má nízkou komunalitu a bylo by vhodné ji vyřadit. Pomocí tohoto pětifaktorového řešení se nám podaří vysvětlit 50,21 % rozptylu. Tato hodnota není až tak vysoká, je tedy zřejmé, že náš model má reálně spíše složitější strukturu.

Zdroje

Institut klinické a experimentální medicíny. (n.d.). Diabetes mellitus. Dostupné z https://www.ikem.cz/cs/diabetes-mellitus/a-4443/

Olorunfemi, B. O., Ogunde, A. O., Almogren, A., Adeniyi, A. E., Ajagbe, S. A., Bharany, S., ... & Hamam, H. (2025). Efficient diagnosis of diabetes mellitus using an improved ensemble method. Scientific Reports, 15(1), 3235.

Teboul, A. (n.d.). Diabetes health indicators dataset. Kaggle. Dostupné z: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.