Identifikace struktur rizikových faktorů diabetu pomocí PCA

Diabetes mellitus představuje chronické metabolické onemocnění charakterizované patologicky zvýšenou hladinou glukózy v krvi. V populaci Pima Indians je tento stav zkoumán pomocí komplexních fyziologických a demografických indikátorů, zahrnujících vedle hladiny glukózy také BMI, krevní tlak, tloušťku kůže, inzulin, věk či reprodukční historii. Rozvoj diabetu je důsledkem narušení homeostázy skrze interakci genetických a fyziologických faktorů. Sledované proměnné nejsou izolované, ale vykazují vysokou vzájemnou provázanost, což ilustruje například vztah mezi BMI a tloušťkou podkožní tkáně. Tato multikolinearita však ztěžuje stanovení individuálního přínosu jednotlivých faktorů k celkovému riziku onemocnění (Knowler et al., 1993).

Metoda

Analýza využívá dataset Pima Indians Diabetes zaměřený na fyziologické indikátory u žen této populace. Původní soubor čítající 768 participantů byl podroben čištění, při němž bylo z důvodu zajištění přesnosti vyřazeno 376 respondentů s neúplnými údaji v klíčových parametrech. Výsledný konzistentní vzorek tvoří 392 participantů s kompletními měřeními všech sledovaných proměnných. Pro zpracování dat v programu Jamovi byla zvolena Analýza hlavních komponent (PCA), která umožňuje transformovat korelující proměnné do nezávislých komponent při zachování maxima původní informace. Před samotnou extrakcí byla data standardizována. Model zahrnuje celkem osm nezávislých proměnných: počet těhotenství, plazmatickou koncentraci glukózy, diastolický krevní tlak, tloušťku kožní řasy, hladinu inzulinu, BMI, skóre genetické zátěže a věk.

Ověření předpokladů pro vícerozměrnou analýzu

Před extrakcí komponent byla ověřena vnitřní struktura dat pomocí dvou diagnostických testů. Kaiser-Meyer-Olkinova (KMO) míra adekvátnosti výběru dosáhla hodnoty 0,73. Vzhledem k tomu, že v odborné literatuře jsou hodnoty nad 0,7 považovány za vysoce vhodné, výsledek potvrzuje dostatečně silné korelace pro smysluplnou redukci dat. Bartlettův test sféricity dále prokázal vysokou statistickou významnost ($p<0,001$). Zamítnutí nulové hypotézy o nezávislosti proměnných definitivně potvrdilo, že

je datový soubor pro aplikaci analýzy hlavních komponent ideálně strukturován (Kaiser, 1974).

Určení počtu komponent

Cílem této fáze analýzy bylo určit optimální počet hlavních komponent, které dostatečně reprezentují variabilitu původních osmi sledovaných proměnných. Pro rozhodování o počtu ponechaných komponent byla využita dvě standardní kritéria:

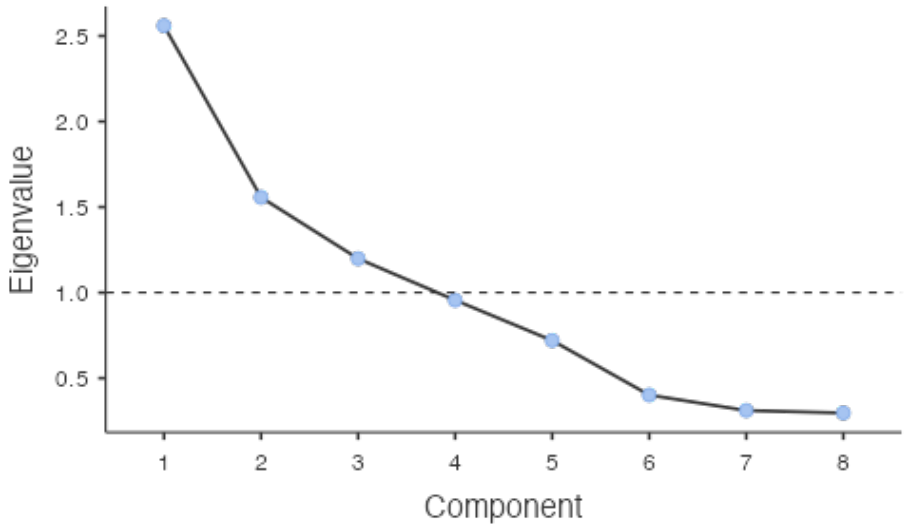
1. **Kaiserovo pravidlo**: Dle tohoto pravidla jsou do modelu zahrnuty pouze ty komponenty, jejichž vlastní číslo je vyšší než 1,0. V našem případě tuto podmínku splnily **tři hlavní komponenty**.
2. **Sutinový graf (Scree Plot)**: Vizuální analýza grafu potvrdila zlom (tzv. „loket“) po třetí komponentě, což naznačuje, že přidávání dalších faktorů by již nepřinášelo významný nárůst vysvětlené informace.

Následující tabulka shrnuje podíl rozptylu, který jednotlivé komponenty pokrývají:

Komponenta	Vlastní číslo	Procento rozptylu	Kumulativní rozptyl
1	2,57	32,0 %	32,0 %
2	1,56	19,5 %	51,5 %
3	1,20	15,0 %	66,5 %
4	0,96	12,0 %	78,5 %

Tabulka 1 Vlastní čísla a vysvětlený rozptyl

Z výsledků vyplývá, že první tři hlavní komponenty dokáží společně vysvětlit **66,5 % celkového rozptylu** původních dat. V kontextu vícerozměrné statistiky a s ohledem na biologickou variabilitu lidského organismu lze tuto hodnotu považovat za velmi uspokojivou, neboť nám umožňuje pracovat s výrazně jednodušším modelem při zachování dvou třetin původní informace.



Interpretace struktury a rotovaná matice zátěží

Pro dosažení tzv. „jednoduché struktury“, která umožňuje jasnou interpretaci nalezených komponent, byla aplikována ortogonální rotace **Varimax**. Tato rotace maximalizuje rozdíly mezi zátěžemi jednotlivých proměnných, čímž nám pomáhá jednoznačně určit, které fyziologické indikátory ke které komponentě náleží.

V tabulce níže jsou uvedeny tzv. zátěže (loadings), které vyjadřují korelaci mezi původní proměnnou a nově vytvořenou komponentou. Pro přehlednost jsou zvýrazněny hodnoty vyšší než 0,5.

Proměnná	Komponenta 1	Komponenta 2	Komponenta 3
BMI	**0,88**	0,06	0,20
Tloušťka kůže	**0,85**	-0,06	0,16
Věk	0,07	**-0,87**	0,19
Počet těhotenství	-0,02	**-0,89**	0,00
Inzulin	0,05	-0,09	**0,86**
Glukóza	0,08	-0,29	**0,80**
Krevní tlak	0,50	-0,46	-0,04
Pedigree Function	0,16	0,10	0,41

Tabulka 2 Rotovaná matice komponent

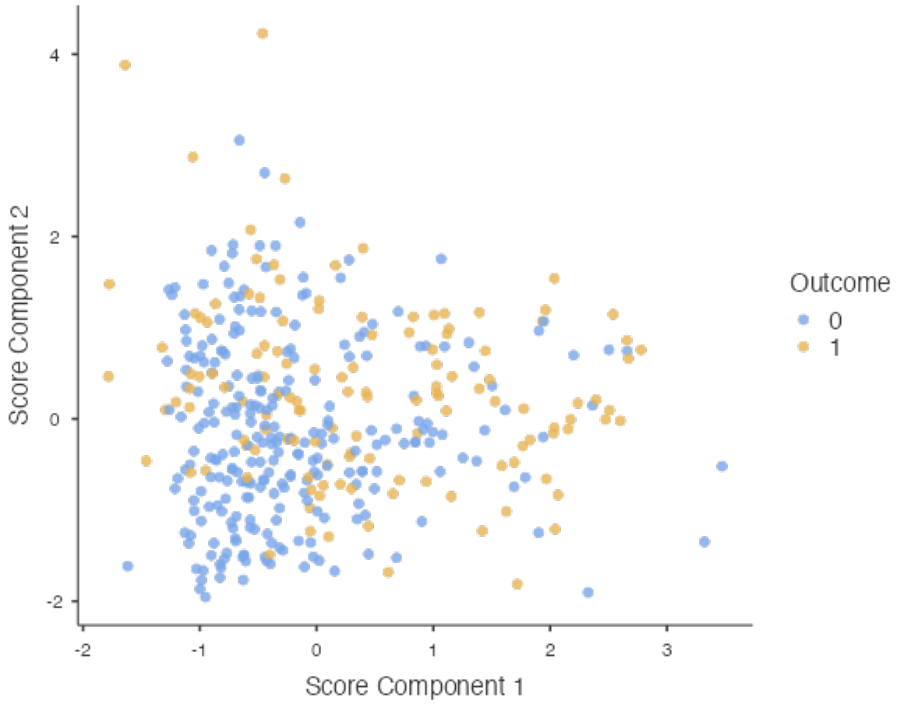
Pojmenování a charakteristika komponent

Na základě statistických zátěží lze identifikovat tři skryté dimenze (faktory), které tvoří pilíře rizikového profilu populace Pima Indians:

1. **Komponenta 1: Antropometrický faktor (Tělesné složení)** Tato komponenta je definována extrémně vysokými zátěžemi indexu BMI a tloušťky kožní řasy. Představuje vnější, fyzickou manifestaci metabolického rizika spojenou s adipozitou (množstvím tuku v těle).
2. **Komponenta 2: Faktor reprodukční historie a stárnutí** Zde se silně projevuje věk respondentek a počet předchozích těhotenství. Tato komponenta zachycuje kumulativní vliv času a biologické zátěže spojené s reprodukčním cyklem, které jsou pro tuto specifickou populaci kritické.
3. **Komponenta 3: Metabolický faktor (Glykemická kontrola)** Tato dimenze sdružuje hladinu glukózy a inzulinu. Představuje „vnitřní" biochemii organismu a schopnost regulovat hladinu cukru, což je primární mechanismus rozvoje diabetu.

Pro ověření praktické relevance nalezených struktur byl vytvořen graf skóre, který zobrazuje jednotlivé respondentky v prostoru první a druhé hlavní komponenty. Body jsou barevně rozlišeny podle přítomnosti diagnózy diabetu (žlutá = pozitivní, modrá = negativní).

Z grafu je patrné, že respondentky s diabetem se koncentrují především v pravé horní části pole. To potvrzuje, že kombinace vysokých hodnot v antropometrickém faktoru (C1 – vysoké BMI) a faktoru věku/reprodukce (C2) představuje nejvýraznější rizikový profil v této populaci.



Závěr

Analýza hlavních komponent úspěšně zredukovala komplexní soubor osmi fyziologických indikátorů na tři stabilní a srozumitelné faktory, které vysvětlují více než dvě třetiny variability v datech. Toto zjištění naznačuje, že rizikové faktory diabetu nepůsobí izolovaně, ale shlukují se do logických celků.

Z praktického hlediska výsledky naznačují, že při screeningu diabetu u populace Pima by se nemělo hledět pouze na izolované hodnoty glykémie, ale je nutné posuzovat pacienta skrze tyto tři dimenze: jeho tělesnou konstituci, biologický věk/anamnézu a aktuální metabolický stav. Tato vícerozměrná perspektiva poskytuje mnohem hlubší vhled do patofyziologie onemocnění než prosté sledování jednotlivých proměnných.

Bibliografie

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*(1)

Knowler, W. C., Saad, M. F., Pettitt, D. J., Nelson, R. G., & Bennett, P. H. (1993). *Determinants of diabetes mellitus in the Pima Indians*. *Diabetes Care, 16*(1), 216–227. https://doi.org/10.2337/diacare.16.1.216

Pelikánová, B. (2025). *Diabetes mellitus – hlavní důvody nárůstu prevalence v populaci* (Bakalářská práce). Univerzita Karlova, Farmaceutická fakulta v Hradci Králové. https://dspace.cuni.cz/bitstream/handle/20.500.11956/198769/130424218.pdf