Predikce absentérství u zaměstnanců[1]

➢ Teoretické ukotvení

Vysoká absence u zaměstnanců může mít nepříznivý dopad na produktivitu organizace. Absence v práci může nastat v důsledku řady faktorů, například věku, zdravotního stavu, nerovnováhy mezi pracovním a osobním životem, nízké motivace, platového ohodnocení a mnoha dalších (Mayfield & Ma, 2020).

V současnosti se řada výzkumů zabývá technikami, které by mohly pochopit a předpovědět absentérství u zaměstnanců. Boehm et al. (2021) zpracovali ve své studii model, který dokáže predikovat absentérství s vysokou přesností. Data byla získána od 740 respondentů a bylo zkoumáno 21 proměnných (např. věk zaměstnance, náklady za dopravu, počet dětí, body mass index a další). Autoři v tomto článku zpracovali data pomocí několika technik strojového učení[2] a zkoumali, která z těchto technik dokáže nejlépe predikovat absentérství u zaměstnanců. Lawrance, Petrides & Guerry (2021) ve studii podobné problematiky tvrdí, že identifikace zaměstnanců, kteří jsou ohroženi vyšší absencí, může posloužit jako prevence a možnost zavedení včasné intervence.

V našem modelu jsme se pokusili predikovat absentérství na základě několika faktorů: věku zaměstnance, peněžních nákladů za dopravu, vzdálenosti bydliště od místa pracoviště, průměrné denní pracovní zátěže, počtu dětí a počtu domácích mazlíčků. Od respondentů byly získány informace, pomocí kterých jsme ověřovali pomocí **logistické regrese**, zda můžeme určit faktory, které dokáží predikovat absentérství u zaměstnanců.

➢ Datový soubor

Datový soubor byl získán na **volně dostupném serveru**[3]. Za účelem aplikace logistické regrese jsme upravili proměnou *Počet hodin absentérství*. Hodnoty této proměnné byly převedeny do binární podoby. Na základě hodnoty mediánu byli respondenti rozděleni do dvou kategorií – jestliže absentovali v zaměstnání více než tři hodiny, byli označeni hodnotou 1, pokud tři hodiny a méně, tak hodnotou 0.

Datový soubor činí 700 respondentů. Testovaní zaměstnanci byli ve věku od 27 do 58 let, přičemž průměrný věk respondenta činil 36 let. Základní vzdělání mělo 79 respondentů (9,57%), středoškolské vzdělání 484 respondentů (58,62%), vysokoškolské vzdělání 273 respondentů (33,05%), postgraduální 11 respondentů (1,33%). V tabulce 1 uvádíme základní popisné charakteristiky.

[1]Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4/zprava.php?id=12.

[2]Strojové učení je podoblastí umělé inteligence, zabývající se algoritmy a technikami, které umožňují počítačovému systému učit se. Strojové učení se značně prolíná s oblastmi statistiky a dobývání znalostí a má široké uplatnění (Matoušek, 2022).

[3] Kaggle – jedná se o platformu, která nabízí uživatelům vyhledávat či publikovat vlastní data. Náš datový soubor (Sharma, 2021) je volně dostupný ke stažení zde: https://www.kaggle.com/datasets/ayushi21095/employee-absenteeism-prediction.

Tabulka 1: ***Popisné charakteristiky***

Prediktor	Průměr	Minimum	Maximum	Směrodatná odchylka
Věk	36,42	27,00	58,00	6,38
Doprava	222,35	118,00	388,00	66,31
Vzdálenost	29,89	5,00	52,00	14,80
Pracovní zátěž	271,80	205,92	378,88	40,02
Dítě	1,02	0,00	4,00	1,11
Domácí mazlíček	0,69	0,00	8,00	1,17

➢ Závisle proměnná a regresory

Zajímáme se o to, které faktory mají vliv na **závisle proměnnou** – **absenci v zaměstnání**. Kladnou odpověď jsme okódovali číslem 1 (jedinec absentoval), zápornou odpověď jsme okódovali číslem 0 (jedinec neabsentoval).

Zkoumali jsme **regresory**, které mohou predikovat **absentérství v zaměstnání**. V našem případě se jedná o tyto regresory:

- *Doprava* (peněžní náklady za dopravu)
- *Vzdálenost* (vzdálenost bydliště od pracoviště v kilometrech)
- *Věk* (věk zaměstnance v letech)
- *Pracovní zátěž* (průměrná denní pracovní zátěž v minutách)
- *Děti* (počet dětí)
- *Domácí mazlíčci* (počet domácích mazlíčků)
- *Vzdělání* (nejvyšší dosažené vzdělání: 1 – základní vzdělání, 2 – středoškolské vzdělání, 3 – vysokoškolské vzdělání, 4 – postgraduální vzdělání)

➢ Výsledky

Pro výpočet **logistické regrese** jsme použili program *Statistica*. V tabulce 2 uvádíme regresní koeficienty, poměr šancí, Waldovu statistiku a p-hodnoty pro jednotlivé prediktory (věk, doprava, vzdálenost, pracovní zátěž, dítě, domácí mazlíček a vzdělání) absentérství u zaměstnanců.

Tabulka 2: ***Výsledky logistické regrese***

Prediktory	Regresní koeficient	Poměr šancí	Waldova statistika	p-hodnota
Konstanta	-1,85	0,16	1,93	0,16
Věk	-0,01	0,99	0,62	0,43
Doprava	**0,01**	**1,01**	**24,57**	**0,00**
Vzdálenost	-0,01	0,99	0,79	0,37
Pracovní zátěž	0,00	1,00	0,97	0,32
Dítě	**0,26**	**1,30**	**10,25**	**0,00**
Domácí mazlíček	**-0,18**	**0,83**	**4,85**	**0,03**

Vzdělání ZŠ	-0,18	0,84	0,03	0,86
Vzdělání SŠ	0,03	1,03	0,00	0,98
Vzdělání VŠ	-0,23	0,79	0,05	0,83

Z výsledků logistické regrese považujeme za statisticky významné proměnné, jejíž **p-hodnota** je menší než hodnota 0,05. Hypotézu o nelineárním vztahu jsme ověřovali pomocí **Waldovy statistiky**. Z těchto postupů vyplývá, že statisticky významné proměnné jsou tyto tři prediktory a jejich interpretace může znít takto:

- *doprava* – šance absence zaměstnance se zvýší 1,01krát, jestliže má zaměstnanec vyšší peněžní náklady za dopravu;
- *dítě* – šance absence zaměstnance se zvýší 1,3krát, jestliže má dítě;
- *domácí mazlíček* – šance absence zaměstnance se sníží 0,83krát, jestliže má domácího mazlíčka.

Kvalitu našeho modelu jsme ověřovali pomocí ukazatelů Cox-Snell R2 a Nagelkerke R2. Tyto hodnoty nám udávají, kolik procent variability jsme schopni vysvětlit pomocí našich regresorů. Přednost dáváme ukazateli Nagelkerke R2, dle kterého náš model vysvětluje 12% variability. Konkrétní hodnoty ukazatelů jsou zobrazeny v tabulce 3. Tabulka 4 znázorňuje předpovězená a pozorovaná data a procento správně predikovaných odpovědí.

Tabulka 3: ***Kvalita modelu I***

	Hodnota	**Hodnota v %**
Cox-Snell R2	0,09	9%
Nagelkerke R2	0,12	12%

Tabulka 4: ***Kvalita modelu II***

	Předpovezená: 1	**Předpovězená: 0**	**Procento správných**
Pozorovaná: 1	157	162	49%
Pozorovaná: 0	98	283	74%

➢ Diskuze

Pomocí logistické regrese jsme ověřili, že náš model dokáže predikovat absentérství zaměstnanců pomocí regresorů doprava, počet dětí a počet domácích mazlíčků. Naše výsledky se shodují například se studií (Erickson & Nichols, 2000) které ukázaly, že rodinné faktory mohou mít významný vliv na neúčast v práci.

Za slabou stránku považujeme nízkou kvalitu modelu, jelikož vysvětluje poměrně malé procento variability. To může být zapříčiněno tím, že nemáme uvedeno dostatečné množství informací od zaměstnanců, které bychom mohli považovat za statisticky významné prediktory, nebo také datový soubor a jeho úprava.

➢ Seznam použité literatury

- Boehm, A., Giering, T., Karg, L. M., & Hermann, T. (2021*). Prediction of Absenteeism at Work using Data Mining Techniques.* International Journal of Advanced Computer Science and Applications, 12(1), 42-49. Dostupné z: https://www.researchgate.net/publication/349868760_Prediction_of_Absenteeism_at_Work_using_Data_Mining_Techniques.
- Erickson, R., & Nichols, L. (2000). *Family influences on absenteeism: Testing an expanded process model.* Journal of Vocational Behavior, 57(2), 246-272. doi: 10.1006/jvbe.2000.1730.
- Lawrance, N., Petrides, G., & Guerry, M.-A. (2021). *Predicting employee absenteeism for cost effective interventions.* Decision Support Systems, 147, 113539. https://doi.org/10.1016/j.dss.2021.113539
- Matoušek, V. (2022). *Strojové učení.* Dostupné z: https://www.kiv.zcu.cz/studies/predmety/uzi/Folie_ZS/Stroj_uceni.pdf.
- Mayfield, M., Mayfield, J., & Ma, K. Q. (2020). *Innovation matters: creative environment, absenteeism, and job satisfaction.* Journal of Organizational Change Management.
- Sharma. A. (2021). *Employee Absenteeism Prediction.* Kaggle. Dostupné z: https://www.kaggle.com/datasets/ayushi21095/employee-absenteeism-prediction.