SUNS – Zadanie 1:

Spracovanie dát, neurónové siete I.

Petra Kirschová

1 NAČÍTANIE A PREDSPRACOVANIE DÁT

Na načítanie a prípravu dát som použila knižnicu `pandas`.

Z datasetu som odstránila:

- Stĺpec *id*
- Riadky, ktoré obsahovali null hodnotu
- Riadky s nezmyselnými údajmi, ako napríklad výška < 130 cm a pod.

V datasete som nahradila reťazce číselnou reprezentáciou:

- Stĺpec *gender*: „woman“ -> 1, „man“ -> 0
- Stĺpce *cholesterol* a *glucose*: „normal“ -> 0, „above normal“ -> 0.5, „well above normal“ -> 1

Do datasetu som pridala nový stĺpec s vypočítanou hodnotou BMI pre každý záznam. Vypočítané hodnoty som skontrolovala a zmazala som ďalšie riadky patriace medzi BMI outliers.

Po úprave dát v datasete zostalo 68 648 záznamov. Vzhľadom na veľkosť celého datasetu (70 000 záznamov) nehrá zmazanie cca 2400 riadkov veľkú rolu.

Následne sa dáta normalizovali, aby sa hodnoty údajov dostali do intervalu <0,1>. Stĺpce, ktoré obsahovali binárne hodnoty 0/1 (*gender, smoke, alco, active* a *cardio*) som po normalizácii konvertovala na integer, pretože normalizáciou sa vygenerovali float hodnoty pre všetky stĺpce. Stĺpec *bmi* som nechala v pôvodnom, nenormalizovanom tvare, keďže sa nepoužíva na trénovanie siete, ale je jej výstupom.

Normalizácia zlepšila výsledky trénovania. Pre porovnanie som viac krát spustila trénovanie binárneho klasifikátora s normalizovanými a nenormalizovanými dátami. Výsledok bol zakaždým lepší v prípade normalizovaných dát. Normalizované dáta tiež dávali oveľa stabilnejšie výsledky – presnosť klasifikátora sa pohybovala okolo 72-73%, pričom pri nenormalizovaných dátach výsledky veľmi kolísali.

Nenormalizované dáta:

```
Binarny klasifikator
Accuracy score:  0.49949031600407745
```

```
Binarny klasifikator
Accuracy score:  0.6286102616377846
```

```
Binarny klasifikator
Accuracy score:  0.5877384592980923
```

Normalizované dáta:

```
Binarny klasifikator
Accuracy score:  0.7291879034998301
```

```
Binarny klasifikator
Accuracy score:  0.7337993301296054
```

```
Binarny klasifikator
Accuracy score:  0.7301587301587301
```

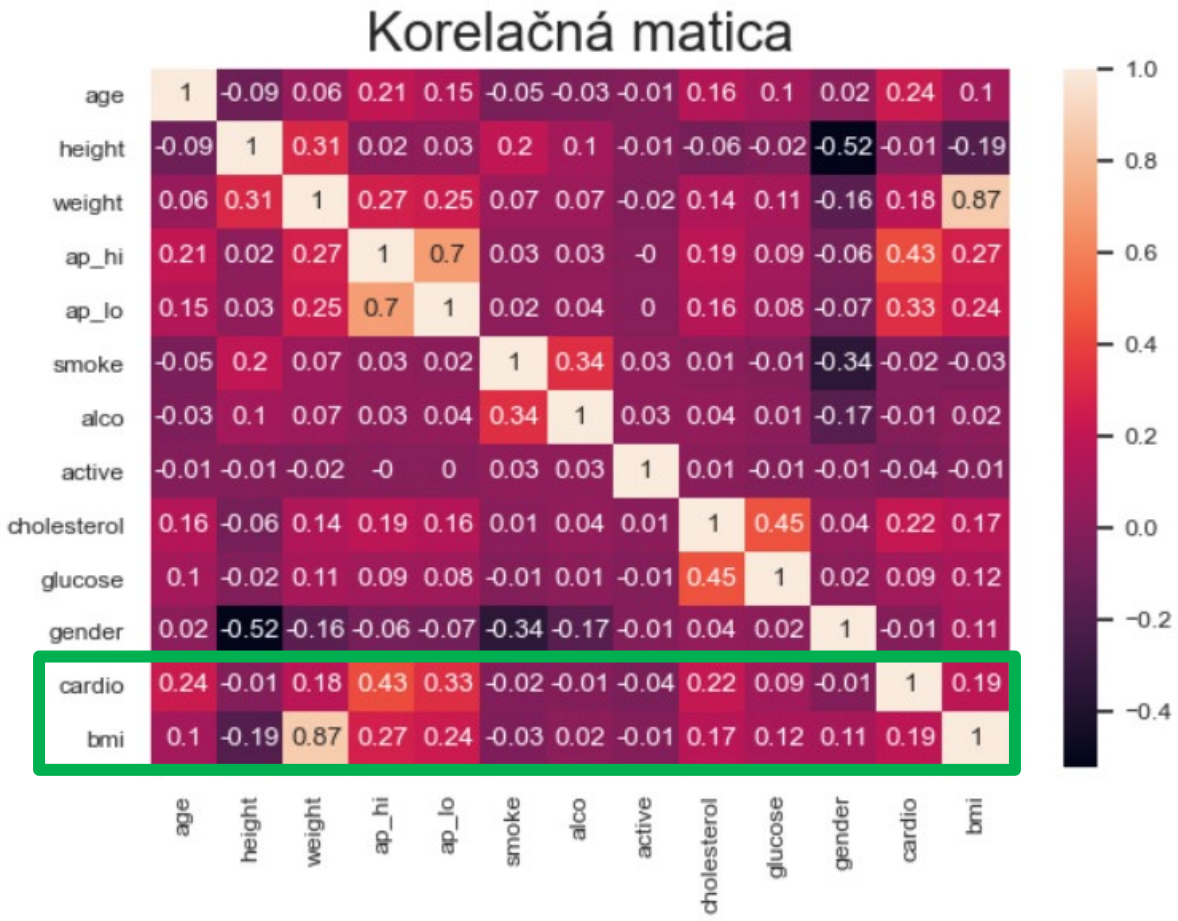
2 ANALÝZA PRÍZNAKOV

Po úprave dát som vykreslila korelačnú maticu pomocou knižníc `seaborn` a `matplotlib`.

Pre **binárny klasifikátor** som sledovala korelácie *cardia* s ostatnými stĺpcami. Z korelačnej matice vyplýva, že *ap_hi* a *ap_lo* (hodnoty krvného tlaku) sú najsilnejšími ukazovateľmi srdcovej choroby. Okrem toho je istá korelácia existencie choroby aj s *age*, *weight* a *cholesterolom*.

Pri **regresore** som sledovala korelačné koeficienty *bmi* s ostatnými stĺpcami. Najvyššia korelácia je s váhou, čo dáva zmysel, keďže BMI sa počíta z váhy a výšky. Ďalšie ukazovatele, ktoré vplývajú na hodnotu BMI, sú hlavne *ap_hi, ap_lo* a tiež *height* (negatívna korelácia), *cholesterol*, *glucose*, *gender* a *cardio*.

Po analýze korelačnej matice som usúdila, že parametre *smoke*, *alco* a *active* majú zanedbateľné korelácie s *cardiom* aj *bmi*, takže ich môžem z datasetu vymazať, keďže pri trénovaní neurónovej siete nebudú smerodajné.



3 BINÁRNY KLASIFIKÁTOR

Na trénovanie klasifikátora som použila knižnicu `sklearn` a na vykreslenie grafov knižnice `seaborn` a `matplotlib`.

3.1 ROZDELENIE DÁT

Najprv som dataset rozdelia na vstupy (X) a výstup (Y) neurónovej siete.

X: Vstupná vrstva neurónovej siete obsahuje 5 neurónov. Vstupnú množinu predstavujú stĺpce *age*, *weight*, *ap_lo*, *ap_hi* a *cholesterol*. Tieto parametre mali najväčší vplyv na prítomnosť srdcovej choroby.

Y: Výstupná vrstva obsahuje 1 neurón, ktorý nadobúda hodnotu 0 v prípade, že osoba nemá chorobu, a 1 ak má chorobu. Výstupnú množinu reprezentuje stĺpec *cardio*.

Ďalej som vstupnú a výstupnú množinu rozdelila na trénovacie a testovacie dáta pomocou funkcie `train_test_split`. Pomer trénovacích a testovacích dát som zvolila 7:3.

3.2 NASTAVENIE KLASIFIKÁTORA

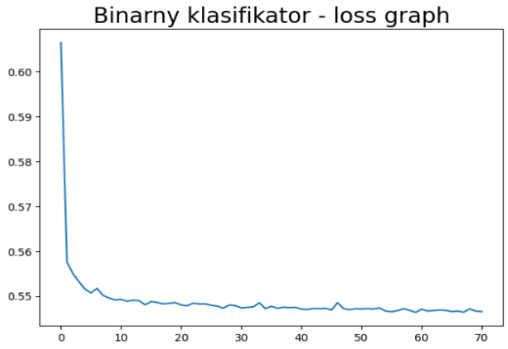
Na trénovanie klasifikátora som použila funkciu `MLPClassifier`. Parametre trénovania som nastavila nasledovne:

```
classifier = MLPClassifier(verbose=True, tol=0.000001, max_iter=300,
                           alpha=0.01, hidden_layer_sizes=(20,))
```

- **tol=0.000001** – tolerancia optimalizácie; zastavovacia podmienka, ak sa počas 10tich po sebe nasledujúcich iterácií skóre nezlepší o túto hodnotu.
- **max_iter=300** – maximálny počet iterácií trénovania, zastavovacia podmienka.
- **alpha=0.01** – penalta
- **hidden_layer_sizes=(20,)** - sieť má 1 skrytú vrstvu, ktorá obsahuje 20 neurónov

3.3 VÝSLEDKY TRÉNOVANIA

Trénovanie bolo ukončené zhruba po 70 iteráciách. Chyba pri trénovaní sa počas prvých 10 iterácií prudko znižovala a ďalej klesala už iba mierne. Na obrázku je zobrazený priebeh chyby počas všetkých iterácií trénovania.

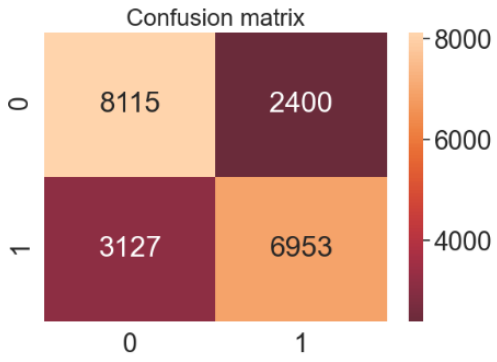


Pre klasifikátor som vygenerovala classification report a confusion matrix. Výsledky klasifikátora boli relatívne dobré, vzhľadom na to, že išlo o reálne dáta. Podarilo sa mi dostať cca 73.2% celkovú presnosť, pričom do kategórie 0 =nemá chorobu bolo správne zatriedených 77.2% (spomedzi všetkých vzoriek, ktoré patrili do kategórie 0) a do kategórie 1 = má chorobu 69% (spomedzi vzoriek patriacich do kat. 1).

********** Binarny klasifikator **********

	precision	recall	f1-score	support
0	0.722	0.772	0.746	10515
1	0.743	0.690	0.716	10080
accuracy			0.732	20595
macro avg	0.733	0.731	0.731	20595
weighted avg	0.732	0.732	0.731	20595

Na confusion matrix je vidno, koľko vzoriek z trénovacej množiny (spolu 20 595 záznamov – 30% z celého datasetu) klasifikátor zatriedil správne a nesprávne. Na diagonále matice je počet správne zatriedených vzoriek do kategórie 0 (nemá chorobu) a 1 (má chorobu). Na opačnej diagonále je počet nesprávne zatriedených vzoriek do jednotlivých kategórií.



4 REGRESOR

Na trénovanie regresora som použila knižnicu `sklearn` a na vykreslenie grafov knižnice `seaborn`, `matplotlib` a `plotly`.

4.1 ROZDELENIE DÁT

Najprv som dataset rozdelia na vstupy (X) a výstup (Y) neurónovej siete.

X: Vstupná vrstva neurónovej siete obsahuje 7 neurónov. Vstupnú množinu predstavujú stĺpce *age, ap_lo, ap_hi, cholesterol, glucose, gender* a *cardio*. Vybrané parametre mali najväčší vplyv na hodnotu *bmi*, s výnimkou *weight* a *height*, keďže z týchto hodnôt sa počítalo BMI, preto sa nemohli použiť.

Y: Výstupná vrstva obsahuje 1 neurón, ktorý vyjadruje hodnotu BMI. Výstupnou množinou je stĺpec *bmi*.

Ďalej som vstupnú a výstupnú množinu rozdelila na trénovacie a testovacie dáta v pomere 7:3, rovnako ako pri binárnom klasifikátore.

4.2 NASTAVENIE REGRESORA

Na trénovanie regresora som použila funkciu `MLPRegressor`. Parametre trénovania zostali totožné s binárnym klasifikátorom:

```
classifier = MLPRegressor(verbose=True, tol=0.000001, max_iter=300,
                          alpha=0.01, hidden_layer_sizes=(20, ))
```

Následne som rovnaké trénovacie a testovacie dáta použila aj pre lineárnu regresiu - funkcia `LinearRegression`.

4.3 VÝSLEDKY TRÉNOVANIA

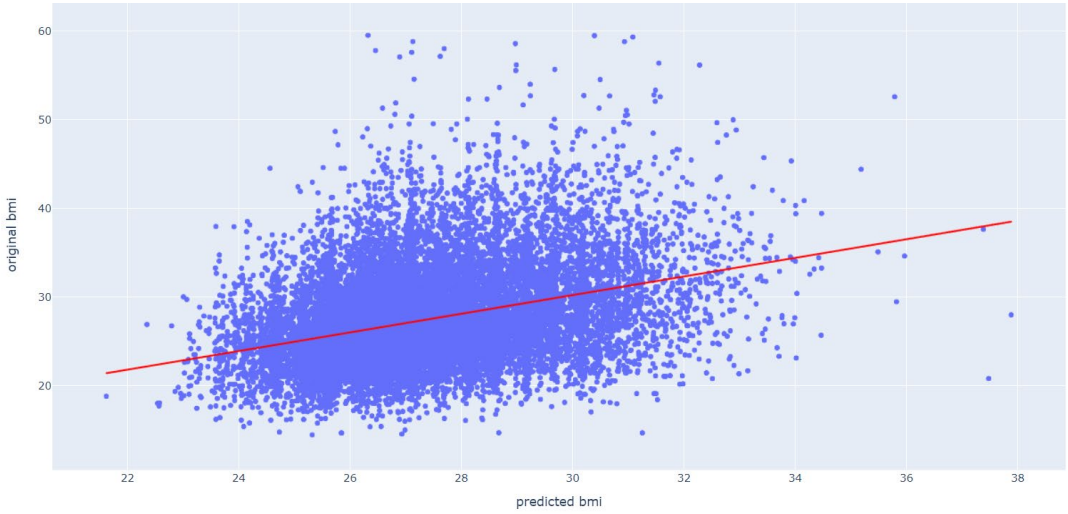
Trénovanie regresora prebiehalo pomerne rýchlo, ukončilo sa približne po 50 iteráciách. Chyba na začiatku prudko klesla a ďalej sa už takmer vôbec nemenila. Na obrázku je vidno, ako prebiehalo trénovanie, resp. aké veľké chyby nastávali pri trénovaní siete počas jednotlivých iterácií.

Po natrénovaní regresora a vypočítaní lineárnej regresie som vypočítala Mean squared error a R^2 score. MLP Regressor dosiahol MSE približne 23.68, teda priemerná odchýlka hodnôt od preloženej regresnej priamky je 23.68. R^2 score, ktoré predstavuje podiel variability, dosiahol cca 12%, čo znamená, že body v okolí priamky sú pomerne dosť rozptýlené. Po testovaní rôznych nastavení siete sa mi nepodarilo presiahnuť hodnotu 12%.

```
*************** MLP Regressor ***************
mse =  23.677270737438423 r2 =  0.12075889689265262
```

Pre regresor som vykreslila Residual plot, ktorý dáva do pomeru predpovedané výstupy neurónovej siete a reálne hodnoty BMI na testovacích dátach. Väčšina bodov sa zhlukuje v strede grafu, kde sa nachádzajú body, ktoré sú veľmi blízko k reálnym hodnotám. Model však nie je dokonalý a teda mnoho bodov ležalo aj ďalej od priamky.

MLP regresor



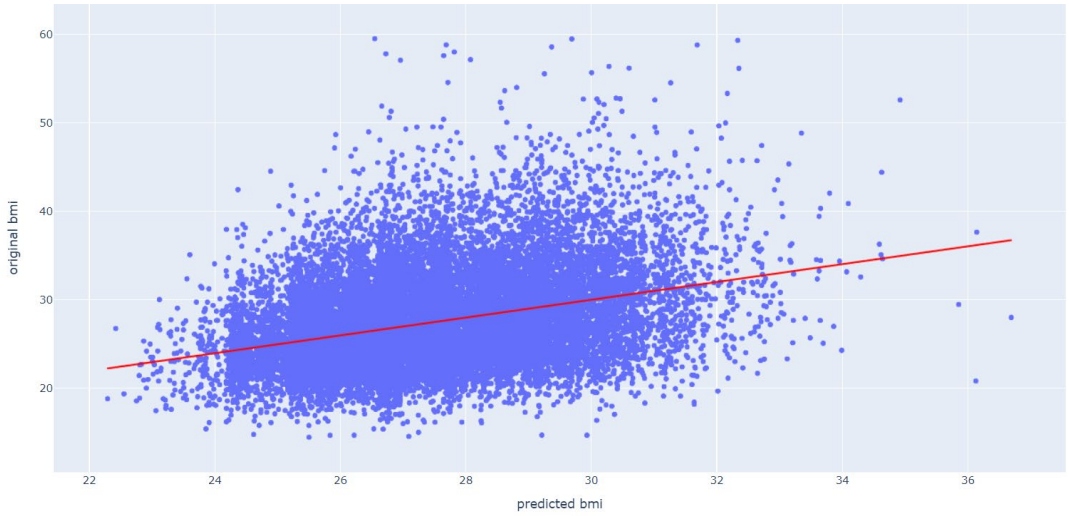
Hodnoty, ktoré model dosiahol, neboli veľmi ideálne, z čoho sa dá usúdiť, že pri trénovaní nemali dané vstupné ukazovatele dostatočnú koreláciu s BMI na to, aby bol regresor natrénovaný lepšie. Váha má najsilnejšiu koreláciu s BMI, ktorá však nemohla byť použitá, nakoľko BMI sa dá pomocou nej jednoducho vypočítať. Sila parametra *weight* sa ukázala, keď som v datasete tento parameter nechala (porovnanie v kapitole 5).

Po natrénovaní regresora som získala výstup z funkcie LinearRegression. Výsledky boli veľmi podobné, líšili sa len na úrovni desatinných miest, dokonca pri trénovaní regresora som získala trochu lepšie výsledky. Regresor mal menšiu MSE chybu, ako pri lineárnej regresii, teda natrénovaná sieť bola trochu úspešnejšia v predpovedaní BMI. R2 score bolo tiež väčšie, aj keď iba o stotinu, teda hodnoty sú trochu menej rozptýlené, ako pri lineárnej regresii. Po viacnásobnom spustení trénovania sa hodnoty regresora a lineárnej regresie stále držali na takmer rovnakej úrovni.

```
*************** Linearna regresia ***************
mse =  23.87246786173828 r2 =  0.11351036994898711
```

Residual graf lineárnej regresie bol na pohľad tiež porovnateľný s Residual grafom natrénovaného regresora.

Lineárna regresia



5 VÝSLEDKY PRE RÔZNE VSTUPNÉ PARAMETRE

5.1 BINÁRNY KLASIFIKÁTOR

Pri trénovaní klasifikátora s pôvodným výberom parametrov som dosiahla celkovú presnosť približne 73%. Pre každú zvolenú kombináciu parametrov som pre porovnanie spustila trénovanie 3 krát.

Pri použití celého datasetu bol výsledok porovnateľný s pôvodným výberom parametrov. Teda napriek tomu, že vstupné parametre medzi sebou obsahovali aj také, ktoré mali s *cardiom* veľmi nízku koreláciu, výsledok tým nebol príliš ovplyvnený.

VSTUP: CELÝ DATASET (okrem cardio)			
pokus	**X**	**Y**	**presnosť (accurracy_score)**
1	age, height, weight, ap_hi, ap_lo, smoke, alco, active, cholesterol, glucose, gender, bmi	cardio	0.727
2	age, height, weight, ap_hi, ap_lo, smoke, alco, active, cholesterol, glucose, gender, bmi	cardio	0.729
3	age, height, weight, ap_hi, ap_lo, smoke, alco, active, cholesterol, glucose, gender, bmi	cardio	0.723

Pri použití iba parametrov, ktoré mali veľmi nízku koreláciu s *cardiom* boli výsledky výrazne horšie. Je to z toho dôvodu, že tieto parametre výsledok takmer vôbec neovplyvňovali a teda na základe týchto ukazovateľov sa neurónová sieť nedokázala naučiť tak, aby dávala spoľahlivé výsledky.

VSTUP: PARAMETRE S NAJNIŽŠÍMI KORELÁCIAMI			
pokus	**X**	**Y**	**presnosť (accurracy_score)**
1	smoke, alco, active, height, gender	cardio	0.535
2	smoke, alco, active, height, gender	cardio	0.534
3	smoke, alco, active, height, gender	cardio	0.518

Keď sa ako vstupy neurónovej siete použili 2 ukazovatele s najvyššími koreláciami, tak výsledky boli relatívne dobré. V porovnaní s použitím viacerých parametrov však bola presnosť trochu menšia. Preto na čo najlepšie natrénovanie siete je lepšie použiť viacero parametrov, aj keď nie všetky z nich majú tak vysokú koreláciu.

VSTUP: PARAMETRE S NAJVYŠŠÍMI KORELÁCIAMI			
pokus	**X**	**Y**	**presnosť (accurracy_score)**
1	ap_hi, ap_lo	cardio	0.713
2	ap_hi, ap_lo	cardio	0.715
3	ap_hi, ap_lo	cardio	0.711

5.2 REGRESOR

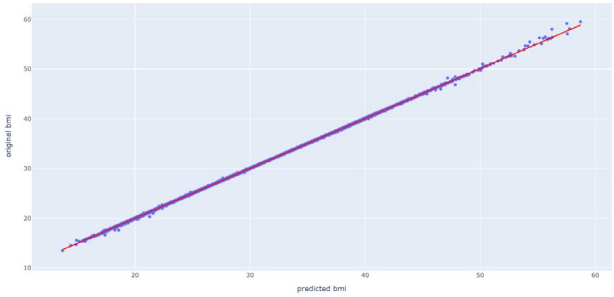
Pri trénovaní regresora som dosiahla MSE cca 23.7 a R2 score 12%. Lineárna regresia mala MSE 23.9 a R^2 11%. Pre každú zvolenú kombináciu parametrov som pre porovnanie spustila trénovanie 3 krát.

Keď bol pri regresii použitý celý dataset okrem *bmi*, medzi parametrami bola aj *weight*, ktorá mala najvyššiu koreláciu s BMI (0.87). Pridanie váhy do vstupnej množiny výrazne ovplyvnilo hodnoty MSE a R^2. Chyba MSE bola zanedbateľná a R^2 score veľmi vysoké, takmer 100%.

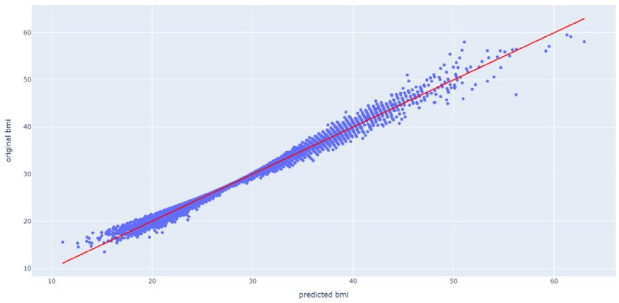
VSTUP: CELÝ DATASET (okrem bmi)						
pokus	**X**	**Y**	**MLP Regressor**		**Linear regression**	
			MSE	**R2**	**MSE**	**R2**
1	age, height, weight, ap_hi, ap_lo, smoke, alco, active, cholesterol, glucose, gender, cardio	bmi	0.0054	0.9998	0.2746	0.9897
2	age, height, weight, ap_hi, ap_lo, smoke, alco, active, cholesterol, glucose, gender, cardio	bmi	0.0312	0.9988	0.2642	0.9901
3	age, height, weight, ap_hi, ap_lo, smoke, alco, active, cholesterol, glucose, gender, cardio	bmi	0.0021	0.9999	0.2497	0.9906

Na grafoch z pokusu 1 je vidno, že regresia bola v tomto prípade veľmi úspešná:

MLP regresor



Lineárna regresia

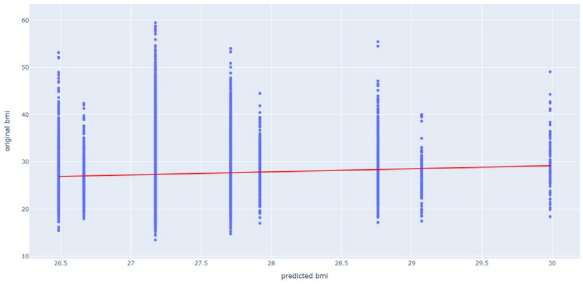


Parametre s nízkou koreláciou dávali veľmi zlé výsledky, v niektorých prípadoch dokonca aj záporné R^2 score. Chyba MSE na druhú stranu ostala podobná, ako pri pôvodnom výbere. Tieto parametre pre regresiu neboli smerodajné, preto výsledné predpovedané BMI je takmer náhodne generované.

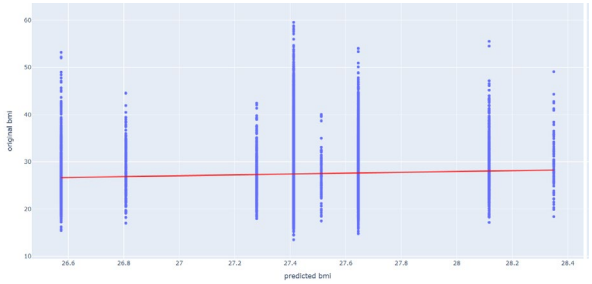
VSTUP: PARAMETRE S NAJNIŽŠÍMI KORELÁCIAMI						
pokus	**X**	**Y**	**MLP Regressor**		**Linear regression**	
			MSE	**R2**	**MSE**	**R2**
1	smoke, alco, active	bmi	26.872	0.0018	26.843	0.0029
2	smoke, alco, active	bmi	26.956	-0.0036	26.798	0.0023
3	smoke, alco, active	bmi	27.4468	-0.00016	27.3826	0.00217

Body v grafe sa nachádzajú v izolovaných priamkach a ich vzdialenosti od regresnej priamky sú väčšinou veľmi veľké, čo tiež značí o nesprávne fungujúcom regresori.

MLP regresor



Lineárna regresia

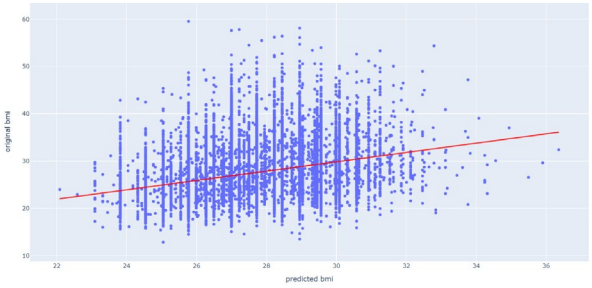


Pri zvolení 2 stĺpcov s najvyššou koreláciou (okrem *weight*) boli výsledky trochu horšie, ako pri pôvodnom výbere, i keď nelíšili sa od nich až tak výrazne. Je však lepšie do vstupnej množiny dať viac veličín, kvôli zlepšeniu kvality trénovania.

VSTUP: PARAMETRE S NAJVYŠŠÍMI KORELÁCIAMI						
pokus	**X**	**Y**	**MLP Regressor**		**Linear regression**	
			MSE	**R2**	**MSE**	**R2**
1	ap_hi, ap_lo	bmi	24.771	0.081	24.998	0.072
2	ap_hi, ap_lo	bmi	25.345	0.074	25.251	0.078
3	ap_hi, ap_lo	bmi	24.761	0.693	24.836	0.066

Grafy vyzerajú podobne, ako pri pôvodnom výbere, avšak sú na nich pozorovateľné pruhy, v ktorých sú body viac centrované, ako v iných miestach.

MLP regresor



Lineárna regresia