SUNS Zadanie 2

Michal Balogh
xbaloghm1@stuba.sk

Október 2025

1 Príprava dát

Príprava a spracovanie dát pre ďalšie úlohy sú v súbore `trees.ipynb`, v sekcii 1. Data Preparation. Dáta som načítal pomocou knižnice `pandas` do `dataframe`. Dataset obsahuje 8741 riadkov a 12 stĺpcov.

Ako prvé som odstránil stĺpec `instant`, keďže je to len identifikátor riadka a nenesie žiadnu informáciu. Taktiež som odstránil stĺpec `date`, keďže by sa zle kódovali pre modely, ktoré budeme používať (365 unikátnych hodnôt). Okrem toho o čase máme informácie zo stĺpcov `month`, `weekday` a `hour`. Rok nepotrebujeme zachovať, pretože všetky údaje sú z jedného roku – 2012.

V ďalšom kroku som skontroloval chýbajúce a duplicitné hodnoty. V datasete chýbalo v stĺpci `holiday` 11 hodnôt (0.13%). Keďže ide o veľmi malú časť dát, tieto riadky som odstránil. Odstránil som aj 8 duplicitných riadkov.

1.1 Odstránenie dát nepatriacich do špecifikovaného rozsahu

Podľa špecifikácie dataset obsahuje údaje s nasledujúcimi rozsahmi pre spojité atribúty:

- temperature: -40, 40
- humidity: 0, 100
- windspeed: 0, 110

A pre diskrétne atribúty:

- month: 1, 12
- hour: 0, 23
- weekday: 0, 6
- holiday: 0, 1
- workingday: 0, 1
- count: ≥ 0

Podľa týchto pravidiel som odstránil 22 riadkov, ktoré mali záporné hodnoty v stĺpci `humidity`.

1.2 Kódovanie kategóriových atribútov

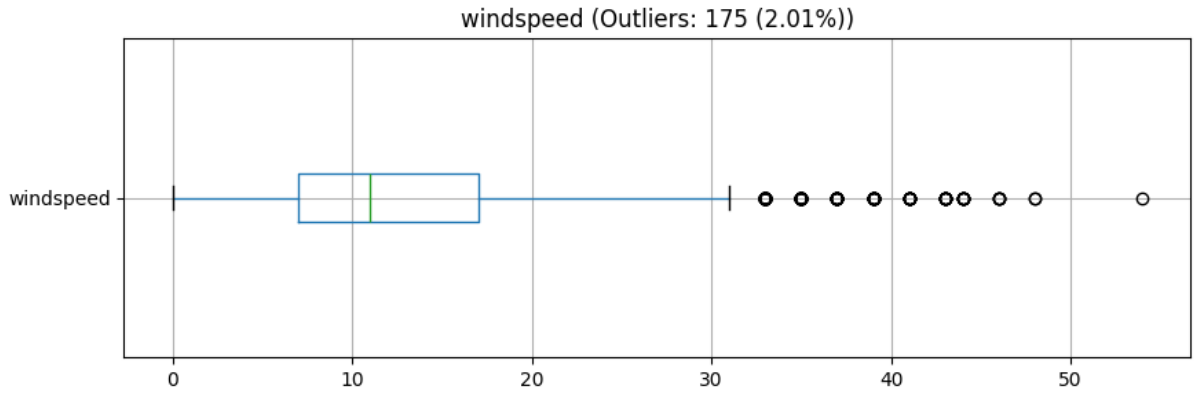
V datasete sme mali 12 stĺpcov, z toho 2 sme odstránili. Zo zvyšných 10 stĺpcov je 9 numerických a len 1 kategóriový – `weather`. Ten som zakódoval cez `label encoding`, keďže má len 4 unikátne hodnoty a poradie medzi nimi dáva zmysel – od najlepšieho po najhoršie počasie.

Kódovanie:

- clear: 0
- cloudy: 1
- light rain/snow: 2
- heavy rain/snow: 3

1.3 Analýza outlierov

Pre analýzu outlierov som použil pravidlo $1.5 \times IQR$ pre spojité atribúty, keďže pre ostatné to nemá zmysel. Outliery má len stĺpec `windspeed`, a to 175 hodnôt (2.01%). Tieto hodnoty som odstránil. Boxplot s outliermi je na obrázku `outliers_windspeed`.

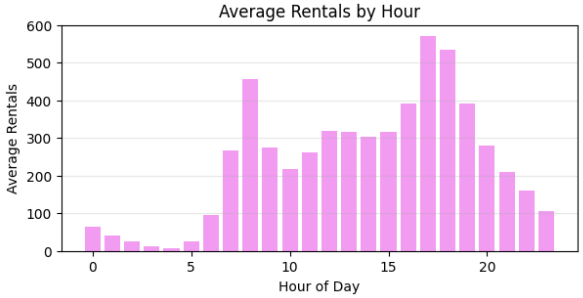


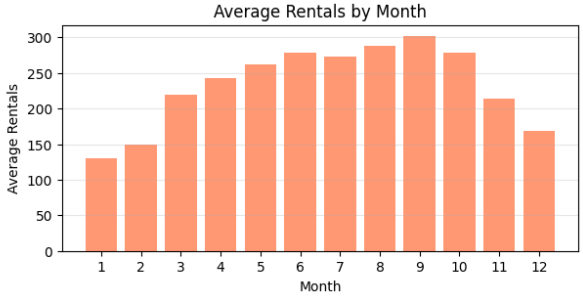
1.4 Finálne rozmery datasetu

Po spracovaní dát mám finálny dataset s rozmermi 8525 riadkov a 10 stĺpcov.

2 EDA

Ako prvé som analyzoval priemerný počet požičaných bicyklov podľa hodiny a mesiaca. Graf je na obrázku `eda_rentals_hour_and_month`.



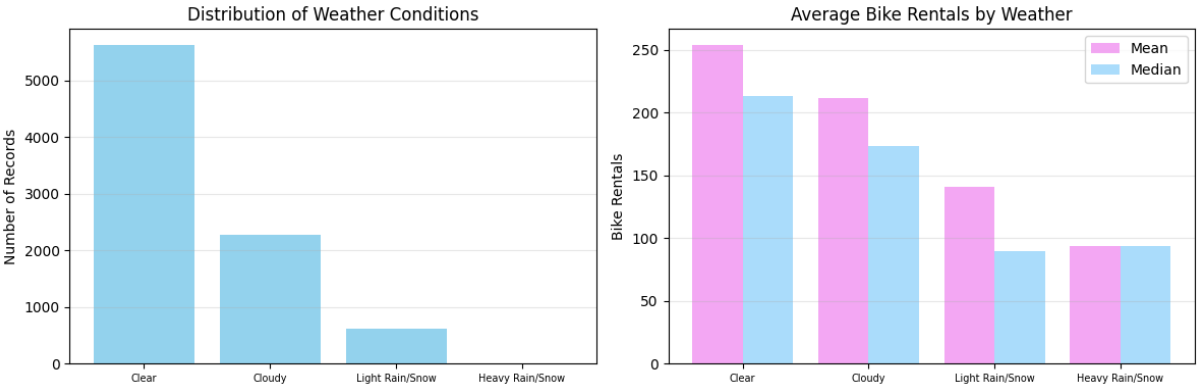


Z grafu vidíme, že bicykle sa viac požičiavajú v teplých mesiacoch (sezóna) – od apríla do septembra. Najvyššie požičiavania sú v mesiaci september. Čo sa týka hodín, najviac si ľudia požičiavajú bicykle ráno okolo 8. hodiny a potom popoludní okolo 17.–18. hodiny. To sú časy, kedy ľudia chodia do práce a z práce.

Ďalší graf zobrazuje priemerný počet požičaných bicyklov v danej hodine v pracovných dňoch (ružová) a cez víkend a sviatky (modrá).

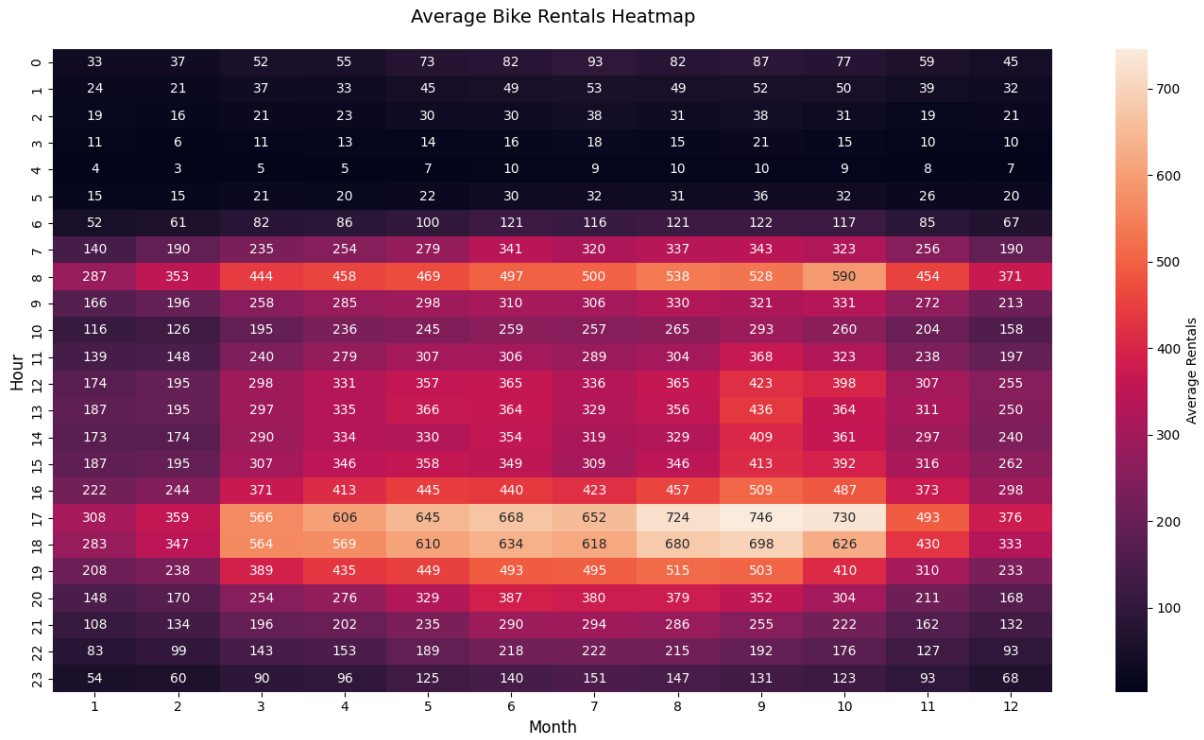
Z grafu vidíme, že v pracovných dňoch je výrazný nárast požičiavaní bicyklov ráno okolo 7.–9. hodiny a potom popoludní okolo 16.–18. hodiny. Naopak, cez víkend a sviatky je požičiavanie bicyklov rozložené rovnomernejšie počas dňa, s vrcholmi okolo 11.–15. hodiny, keď ľudia využívajú bicykle na rekreáciu.

Na grafe `eda_weather` je zobrazená distribúcia počasia v datasete. Na grafe vľavo je priemer a medián počtu požičiavaní bicyklov pre jednotlivé kategórie počasia.



Najviac bicyklov sa požičiava počas jasného (clear) a počas oblačného (cloudy) počasia.

Graf `eda_heatmap` zobrazuje priemerný počet požičaných bicyklov podľa hodiny a mesiaca v heatmape.



Na heatmape je najjasnejší vrchol v mesiacoch august, september a október okolo 17.–18. hodiny. Ďalšie vrcholy sú v mesiacoch marec až júl, tiež okolo 17.–18. hodiny. Okrem toho ešte v marci okolo ôsmej hodiny ráno.

Zhrnutie vrcholov požičiavaní bicyklov počas pracovných a nepracovných dní je v tabuľke `eda_rentals_table`.

Peak Hours Analysis

Category	Peak Hour	Avg Rentals
All	17	571.7
Workday	17	654.4
Weekend/Holiday	12	460.7

3 Rozdelenie dát na trénovacie a testovacie

Dáta som rozdelil na trénovacie a testovacie v pomere 8:2 pomocou funkcie `train_test_split` z knižnice `sklearn.model_selection`. Dáta som škáloval pomocou `StandardScaler`.

4 Tréning modelov

Tréning modelov je v súbore `trees.ipynb`, v sekcii 4. Model Training. Použil som tri modely: Decision Tree Regressor, Random Forest Regressor a Support Vector Machine z knižnice `sklearn`.

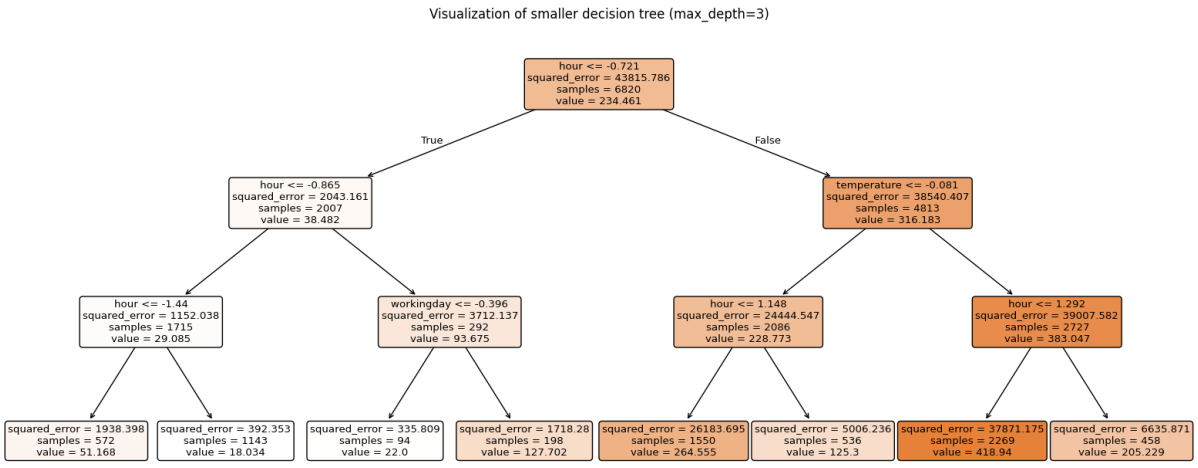
4.1 Decision Tree

Na nájdenie najlepšej hĺbky rozhodovacieho stromu som postupne cez cyklus skúšal hodnoty od 1 do 100. Najlepšiu hodnotu som našiel pri hĺbke 10. Model dosiahol na testovacích dátach $R^2 = 0.889$.

Tabuľka s výsledkami pre hĺbku stromu 1 až 10 je na obrázku `tree_max_depth_table`.

Tree max_depth=	1	2	3	4	5	6	7	8	9	10
R2 score	0.353	0.438	0.525	0.603	0.649	0.775	0.848	0.868	0.885	0.889
Number of nodes	3	7	15	31	63	127	253	489	911	1611
Number of leaves	2	4	8	16	32	64	127	245	456	806
Height	1	2	3	4	5	6	7	8	9	10

Keďže hĺbka 10 sa už zle vizualizuje, zvolil som pre vizualizáciu stromu hĺbku 3. Vizualizácia rozhodovacieho stromu je na obrázku `tree_viz`.



Na strome môžeme vidieť, že najdôležitejší atribút pre rozhodovanie je atribút `hour`. Je v koreni stromu a zároveň v ďalších štyroch uzloch – dokopy 5 zo 7 rozhodovacích uzlov. Následne je dôležitý atribút `temperature` a `workingday`.

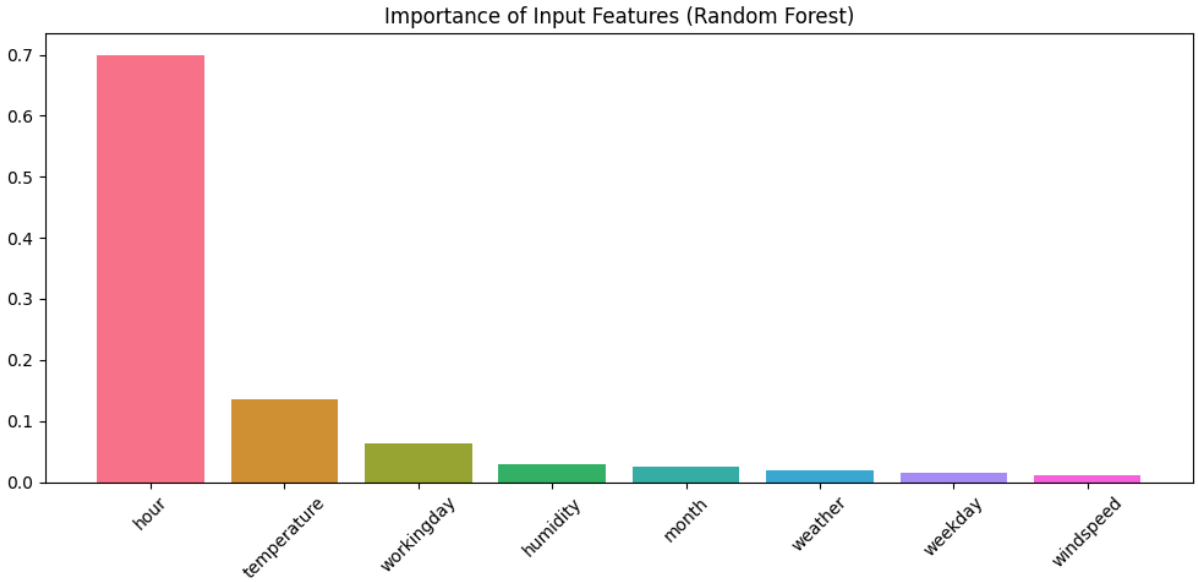
Tento strom s hĺbkou 3 dosiahol na testovacích dátach $R^2 = 0.525$.

4.2 Random Forest

Pre Random Forest som skúšal rôzny počet stromov v lese (`n_estimators`) a rozhodol som sa pre hodnotu 10, ktorá síce nedosiahla najlepšie výsledky, ale bola dostatočne rýchla a zväčšovaním počtu stromov sa už výsledky veľmi nezlepšovali.

Model dosiahol na testovacích dátach s počtom stromov v lese 10 $R^2 = 0.925$. Pre počet stromov 100 dosiahol model $R^2 = 0.934$, čo je zlepšenie len o 0.009 oproti 10 stromom.

Príznaky, podľa ktorých sa rozhodoval Random Forest, sú zobrazené na obrázku `importance_of_input_features`.



Potvrdilo sa, že najdôležitejší atribút je `hour`, následne `temperature` a `workingday`. Atribút `hour` poskytuje približne 70% dôležitosti v rozhodovaní modelu, následne `temperature` približne 13% a `workingday` okolo 6%. Zvyšné atribúty už majú podobne malú dôležitosť.

4.3 Support Vector Machine (SVM)

Pre Support Vector Machine som skúšal rôzne jadrá (`kernel`) a najlepšie výsledky som dosiahol s jadrom `rbf` (radial basis function). Parameter C (regularizácia) som nastavil na hodnotu 100, ktorá dosiahla lepšie výsledky ako nižšie hodnoty.

Model dosiahol R^2 skóre 0.554.

4.4 Porovnanie a vyhodnotenie modelov

Porovnanie všetkých troch modelov (Decision Tree, Random Forest a SVM) je uvedené v tabuľke nižšie. Modely boli vyhodnotené pomocou metrík R^2, RMSE a MSE na trénovacích aj testovacích dátach.

Model Comparison

Model	Train R2	Test R2	Train RMSE	Test RMSE	Train MSE	Test MSE
Decision Tree	0.950	0.889	46.80	69.17	2189.81	4784.76
Random Forest	0.989	0.925	22.22	57.00	493.87	3249.53
SVM	0.579	0.554	135.81	138.77	18443.19	19258.26

Random Forest dosiahol najlepšie výsledky zo všetkých troch modelov:

- Test R^2 skóre: 0.925, čo znamená, že model vysvetľuje 92.5% variability v dátach
- Test RMSE: 57.00, čo predstavuje priemernú odchýlku predikcie
- Vysoké trénovacie R^2 (0.989) naznačuje mierne pretrénovanie (overfitting), ale rozdiel oproti testovaciemu R^2 nie je dramatický

Decision Tree s optimálnou hĺbkou 10 dosiahol:

- Test R^2 skóre: 0.889, čo je o 0.036 horšie ako Random Forest
- Test RMSE: 69.17
- Výrazný rozdiel medzi trénovacím (0.950) a testovacím R^2 naznačuje mierne pretrénovanie

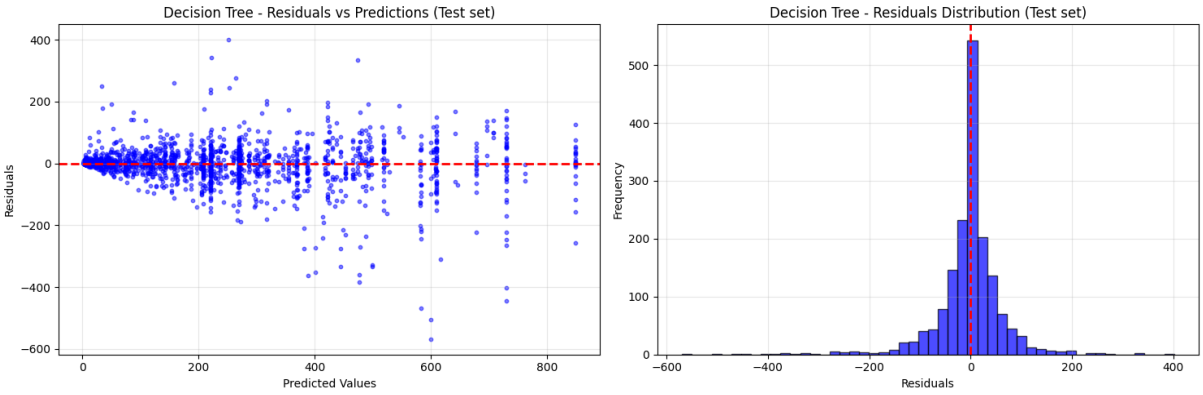
Support Vector Machine (SVM) dosiahol najhoršie výsledky:

- Test R^2 skóre: 0.554, výrazne nižšie ako stromové modely
- Test RMSE: 138.77, viac ako dvojnásobne vyššie ako Random Forest
- Podobné výsledky na trénovacích aj testovacích dátach (0.579 vs. 0.554) naznačujú, že model nie je pretrénovaný, ale len nedokáže dobre zachytiť vzťahy medzi dátami s daným nastavením hyperparametrov

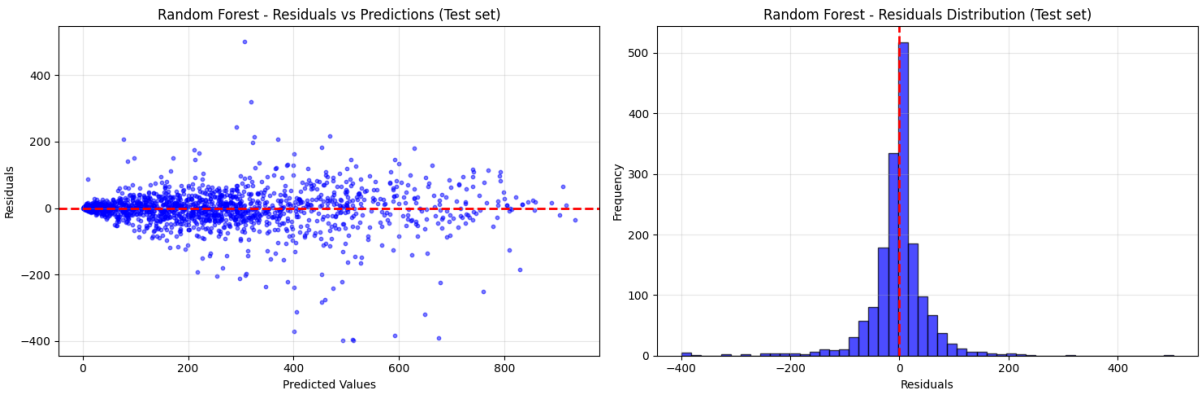
Analýza reziduálov

Grafy reziduálov (rozdiely medzi skutočnými a predikovanými hodnotami) pre všetky tri modely sú zobrazené nižšie.

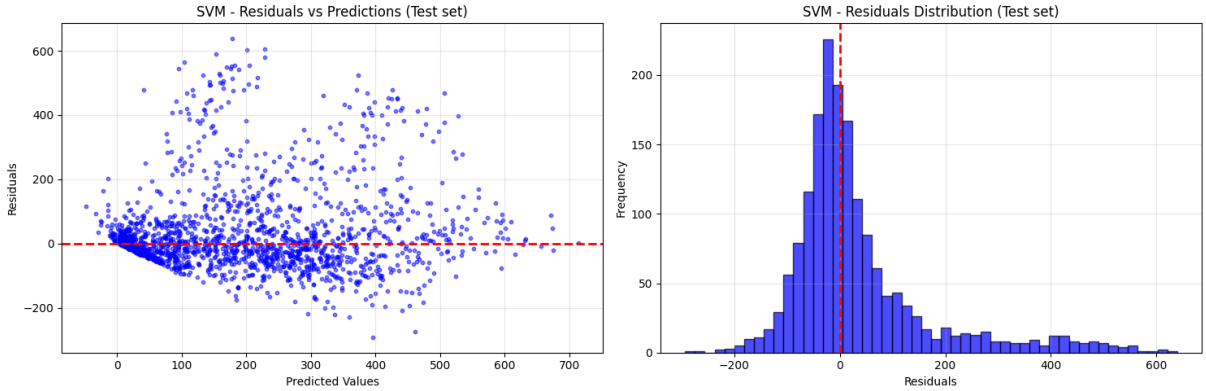
Decision Tree:



Random Forest:



SVM:



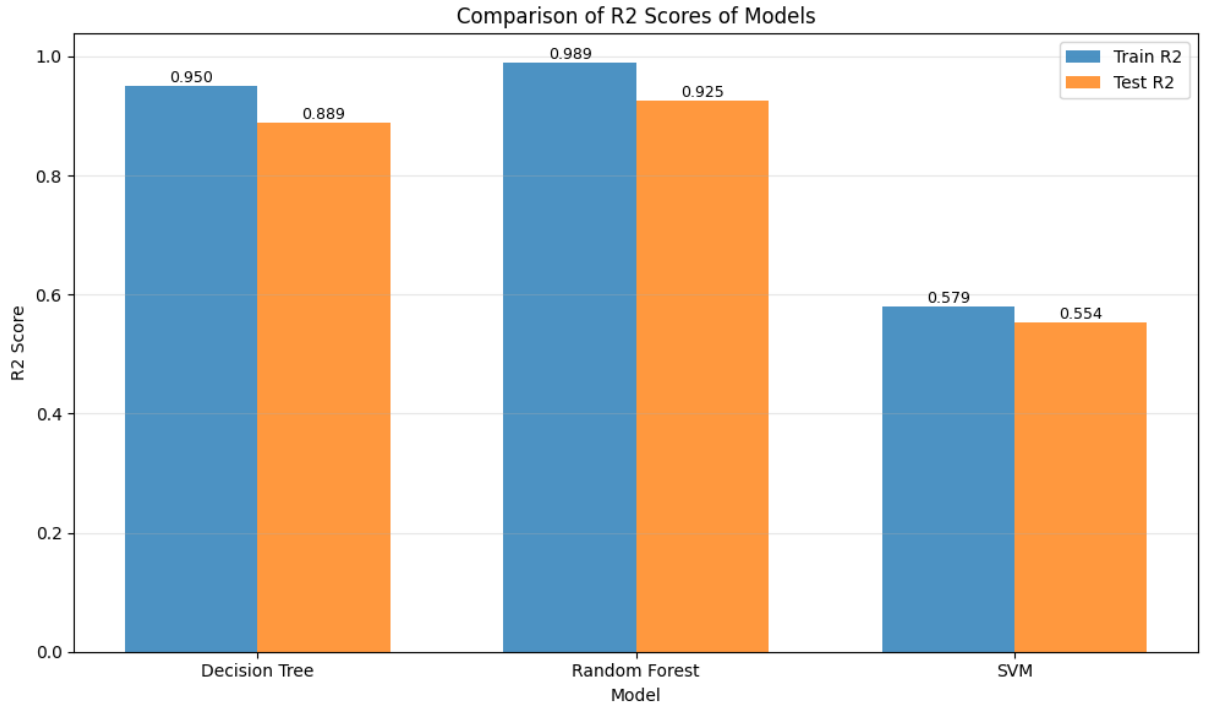
Z grafov reziduálov možno pozorovať, že Random Forest má najmenšie rozptýlenie reziduálov a najlepšie sa približuje k ideálnemu stavu (reziduály okolo nuly). Distribúcia reziduálov je symetrická okolo nuly (približne normálne rozdelená), čo je znak dobrého modelu.

Decision Tree má o niečo väčšie rozptýlenie, ale stále prijateľné. Distribúcia reziduálov je mierne zošikmená doľava.

SVM už má značné chyby s veľkým rozptýlením reziduálov. V grafe s distribúciou reziduálov sú dáta zošikmené sprava, čo znamená, že model má tendenciu podhodnocovať niektoré vysoké hodnoty (predpovede sú príliš nízke oproti realite). Inak povedané – model nevie dobre zachytiť extrémne vysoké hodnoty cieľovej premennej.

Porovnanie R^2 skóre na množine train a test

Vizuálne porovnanie R^2 skóre pre všetky modely pre trénovacie a testovacie dáta je na grafe `r2_comparison`.

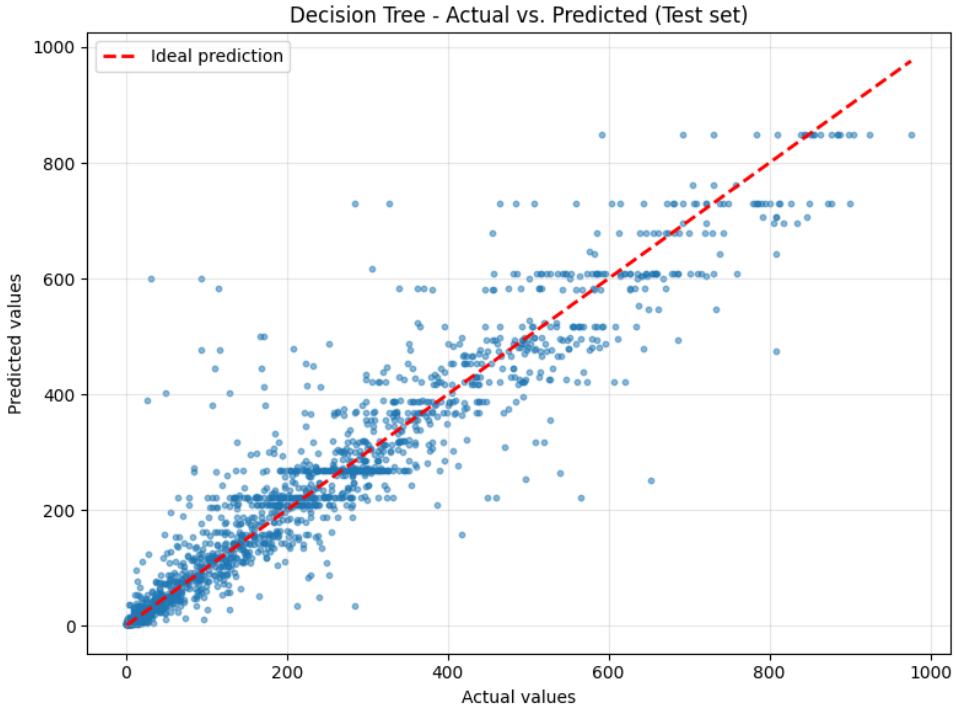


Z grafu vidno, že medzi trénovacími a testovacími dátami nie je až taký veľký rozdiel. Z toho vyplýva, že modely nie sú pretrénované.

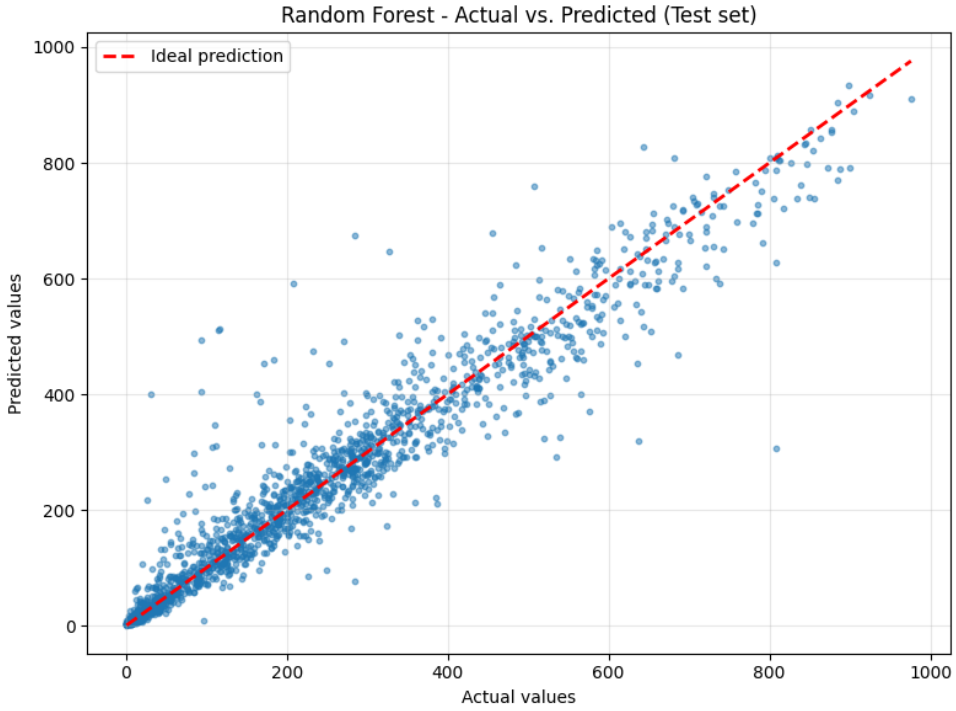
Predikcie a skutočné hodnoty

Grafy porovnania predikovaných a skutočných hodnôt pre testovací súbor pre všetky modely sú na grafoch nižšie.

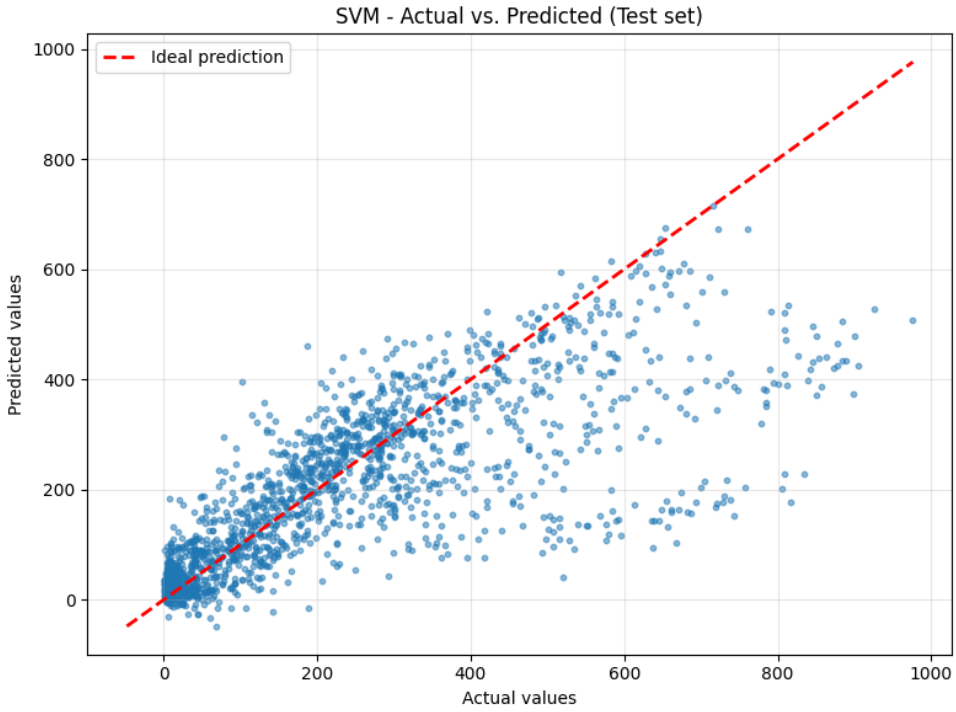
Decision Tree:



Random Forest:



SVM:



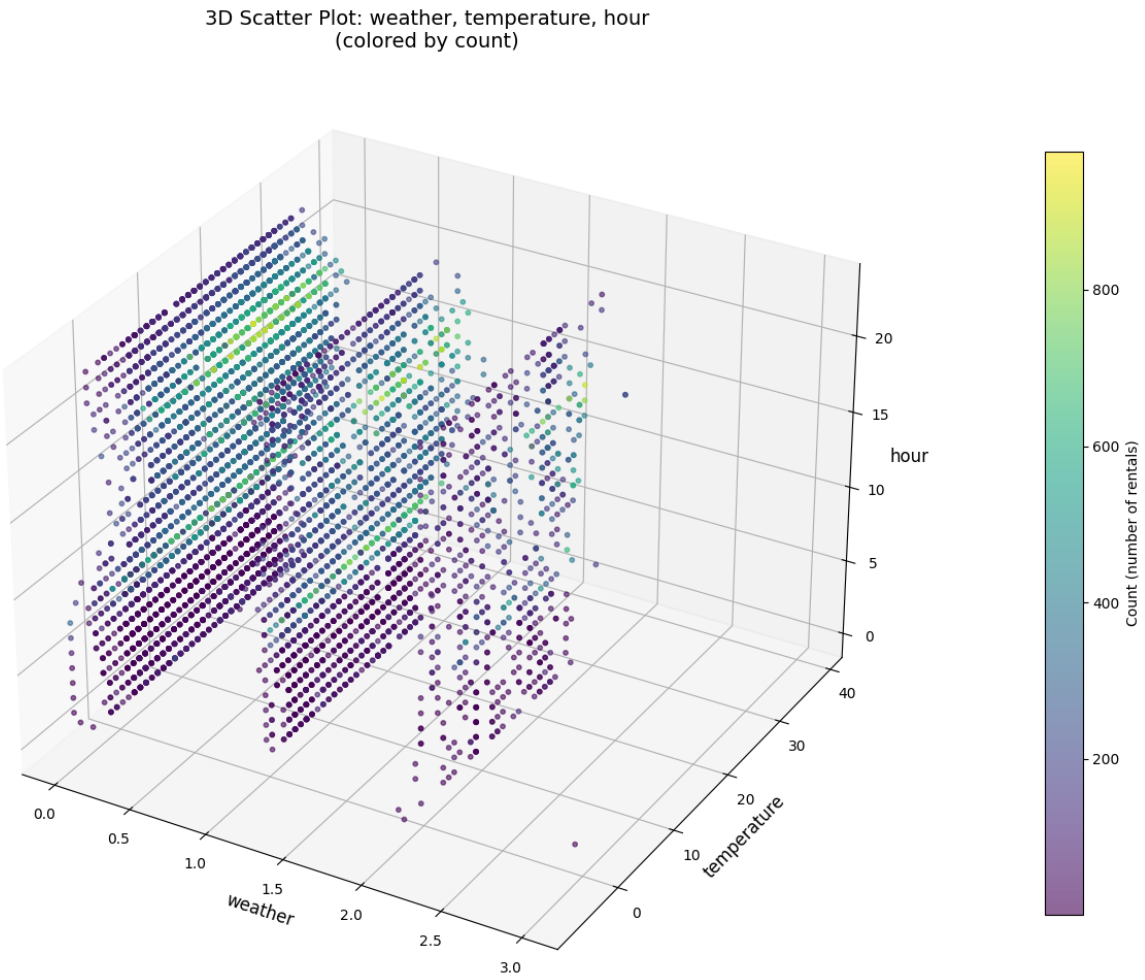
Ideálne predikcie by mali ležať na červenej čiare ($y = x$). Väčšina bodov sa drží blízko diagonály, najmä pri nižších hodnotách, takže model sa celkom dobre trafí pre menšie reálne hodnoty. Pri SVM pri väčších hodnotách (napr. nad 400–500) sa body začínajú rozptyľovať pod čiaru, čo znova potvrdzuje, že model podhodnocuje vyššie hodnoty.

5 Redukcia dimenzie

V tejto časti som vizualizoval dáta v 3D priestore pomocou dvoch metód: 3D scatter plotu, kde som vybral tri príznaky, a PCA (Principal Component Analysis).

5.1 3D Scatter Plot

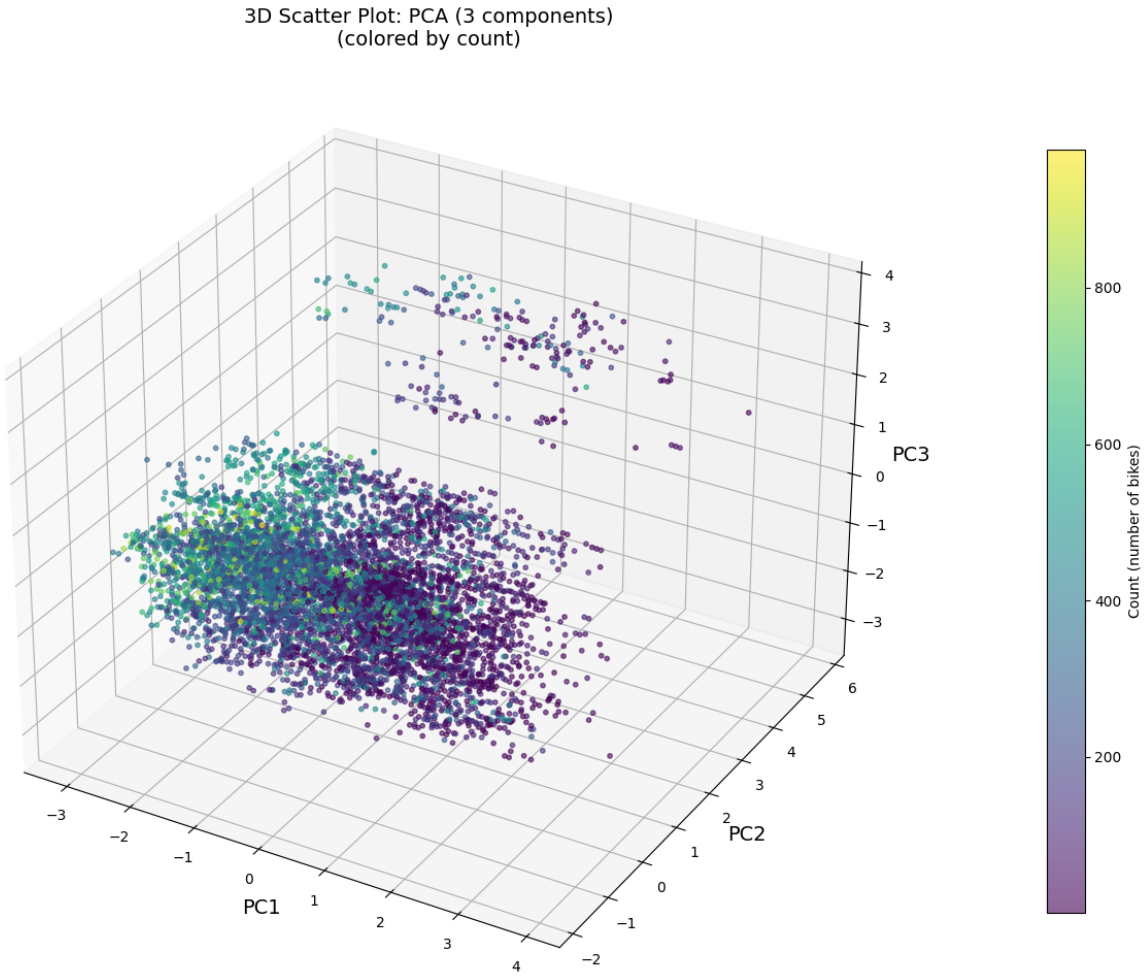
Ako tri príznaky, ktoré som vyniesol do 3D grafu, som zvolil `hour`, `temperature` a `weather`. Prvé dva sú najdôležitejšie príznaky podľa analýzy dôležitosti príznakov z modelu Random Forest. Tretím príznakom je `weather`, ktorý som zvolil preto, lebo sa jednoduchšie interpretuje. Atribút `weather` má 4 unikátne hodnoty, ktoré sú reprezentované číslami 0–3 po label encodingu.



Atribút `weather` nám rozdelil dáta na pekné štyri zhluky. V poslednom zhluku – `heavy rain/snow` (3) je len jedna bodka – počas veľmi nepriaznivého počasia je požičiavanie bicyklov veľmi nízke. Na druhej strane, pri počasí `clear` (0) je najväčší počet bodov a z ďalších dvoch osí môžeme vyčítať, že najviac požičiavaní bicyklov je medzi 14.–16. hodinou a pri teplote okolo 20–30 °C.

5.2 PCA

Po aplikácii PCA na škálované trénovacie dáta som získal graf na obrázku `pca_3d`.



Väčšina bodov je v hustej oblasti, čo znamená, že tieto dáta majú podobné vlastnosti. Nie sú zrejmé žiadne jasné klastre. Bodky s vysokými hodnotami `count` (žlté) sa sústreďujú relatívne blízko v zhlukoch grafu, čo môže indikovať, že existujú kombinácie podmienok (počasie, deň v týždni, hodina...), ktoré vedú k vysokej aktivite bicyklov. Outliery mimo hlavný zhluk môžu predstavovať špecifické situácie, ako napríklad extrémne počasie alebo sviatky.

Po redukcii dimenzie na tri hlavné komponenty PCA som analyzoval váhy (loadings) pôvodných premenných na týchto komponentoch:

Komponenta	**Exp var**	**Loadings**	**Interpretácia**
PC1	19.2%	+hum, +weather, −temp, −wind	Počasie (nepriaznivé vs. priaznivé)
PC2	14.9%	+holiday, −workingday, −weekday	Typ dňa (víkend vs. pracovný deň)
PC3	13.2%	+month, +temp, −wind	Sezóna (leto vs. zima)
Celkovo	47.3% var	–	Takmer polovica informácií v dátach

Tabuľka 1: Hlavné komponenty PCA a ich interpretácia

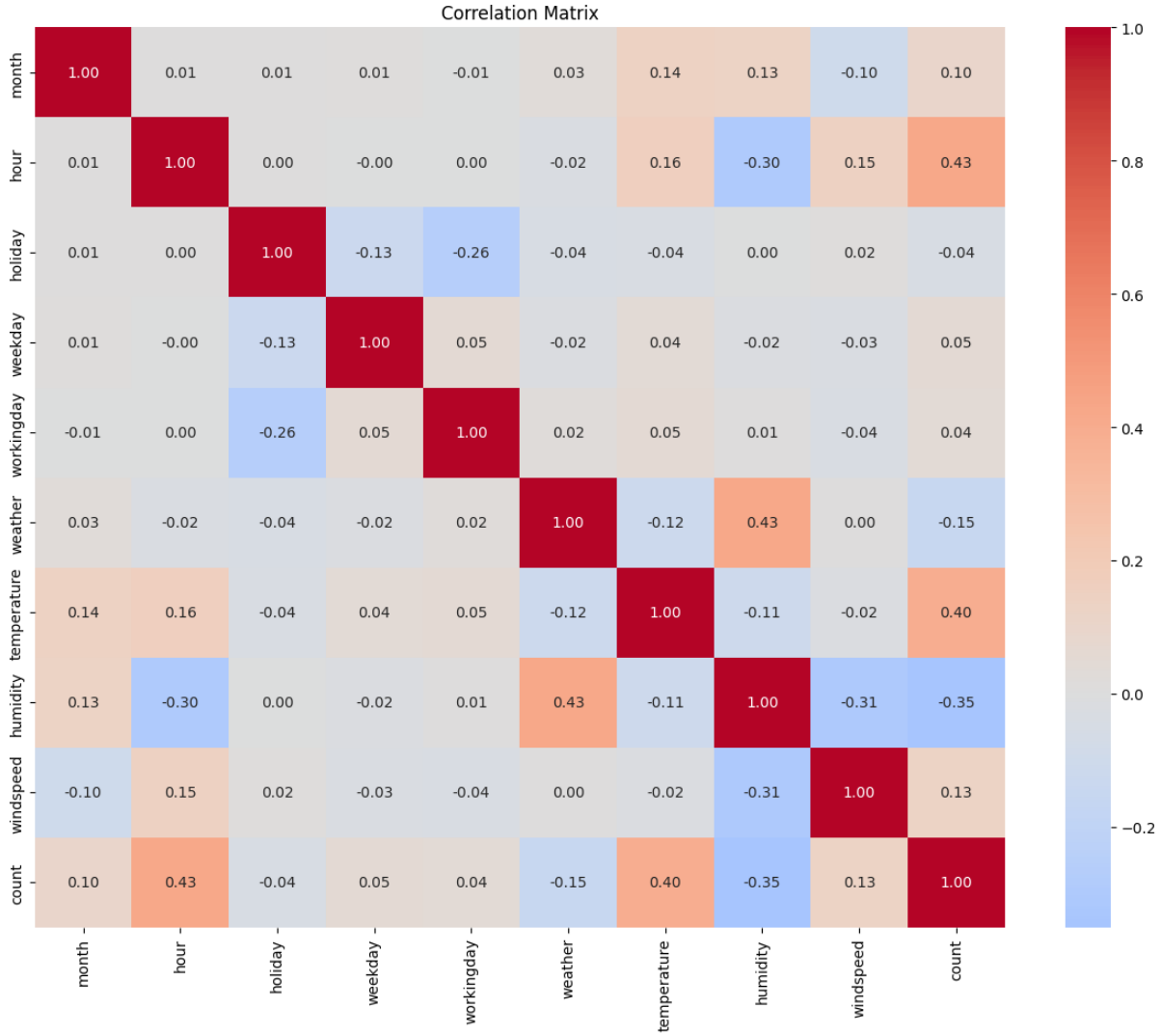
6 Trénovanie na podmnožine príznakov

V tejto časti som trénoval najlepší model – **Random Forest** na zmenšenej množine príznakov. Množinu príznakov som zmenšil tromi metódami:

- podľa korelačnej matice,
- podľa dôležitosti príznakov z Random Forest,
- pomocou PCA.

6.1 Korelačná matica

Korelačná matica príznakov je na obrázku `corr_mat`.



Obr. 1: Korelačná matica príznakov.

Príznaky, ktoré korelovali s cieľovou premennou `count` viac ako stanovený prah 0.1, som vybral ako nové príznaky pre trénovanie modelu. Týchto príznakov bolo 5: `hour`, `temperature`, `humidity`, `weather` a `windspeed`.

Model dosiahol na testovacích dátach $R^2 = 0.694$. Test RMSE sa zhoršilo na 114.86.

6.2 Dôležitosť príznakov z Random Forest

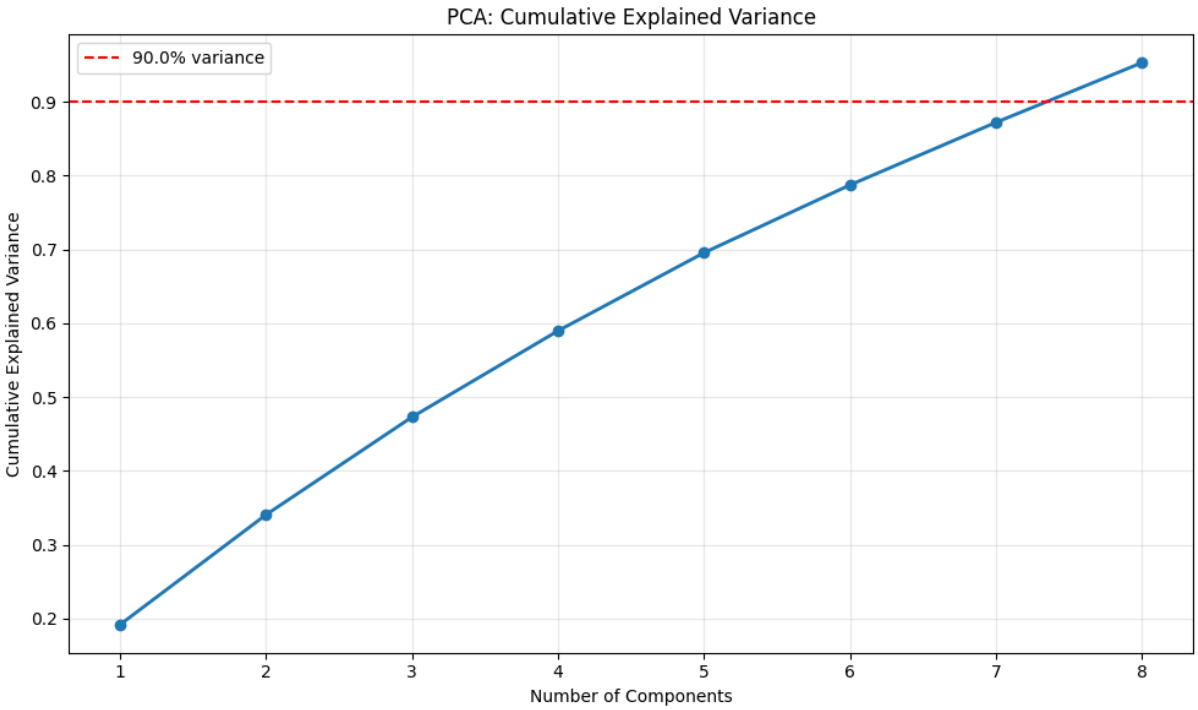
Graf dôležitosti príznakov z Random Forest už bol v sekcii 4.2, na obrázku `importance_of_input_featu`

Kumulatívnym súčtom som chcel pokryť aspoň 90 % dôležitosti príznakov. To som dosiahol len s tromi príznakmi: `hour`, `temperature` a `workingday`. To znamená, že tieto tri príznaky zodpovedajú za 90 % rozhodovania modelu.

Model dosiahol na testovacích dátach $R^2 = 0.839$. Test RMSE sa zhoršilo na 83.30.

6.3 PCA

Pre PCA som zvolil prah 90 % vysvetlenej variability. Túto hodnotu som dosiahol s ôsmimi komponentmi. Osem komponentov pokrylo 95.3 % variability dát. Kumulatívny súčet vysvetlenej variability je na obrázku `cumsum_expl_var`.



Obr. 2: Kumulatívny súčet vysvetlenej variability PCA.

Model dosiahol na testovacích dátach $R^2 = 0.583$. Test RMSE sa zhoršilo na 134.14.

6.4 Porovnanie výsledkov na zmenšenej množine príznakov

Porovnanie výsledkov modelu Random Forest na zmenšenej množine príznakov je v tabuľke `feature_selection_methods_table`.

Feature Selection Methods Comparison

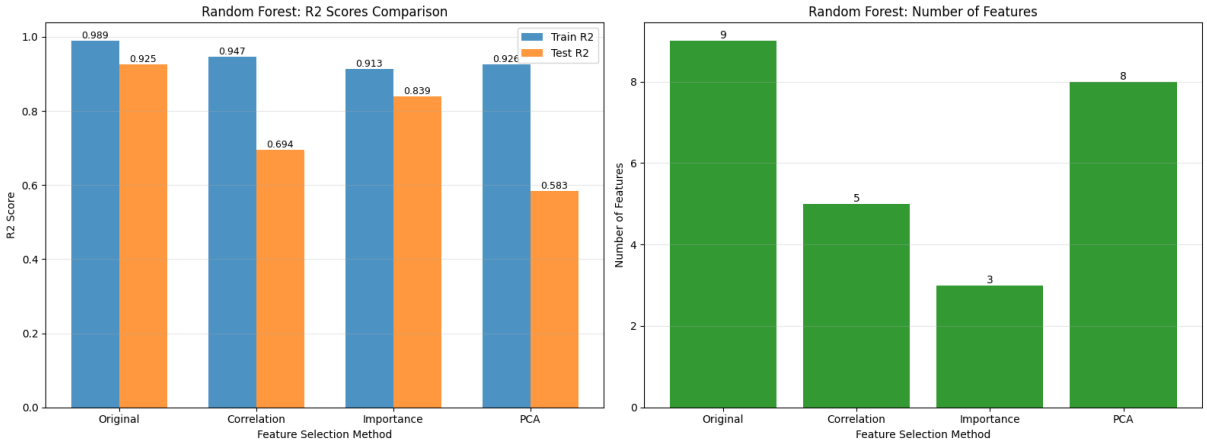
Method	Features	Train R2	Test R2	Train RMSE	Test RMSE	Train MSE	Test MSE
Random Forest (Original)	9	0.989	0.925	22.22	57.00	493.87	3249.53
Random Forest (Correl)	5	0.947	0.694	48.22	114.86	2325.59	13192.56
Random Forest (Important)	3	0.913	0.839	61.71	83.30	3808.39	6938.48
Random Forest (PCA)	8	0.926	0.583	56.89	134.14	3236.76	17992.31

Obr. 3: Porovnanie výsledkov Random Forest na zmenšenej množine príznakov.

Najlepšie výsledky boli dosiahnuté na pôvodnej množine príznakov. Zmenšenie množiny príznakov spôsobilo zhoršenie výsledkov vo všetkých prípadoch. Najmenej sa výsledky zhoršili pri výbere príznakov podľa dôležitosti príznakov z Random Forest. Metrika R^2 sa zhoršila o 0.086 a RMSE sa zhoršilo o 26.30.

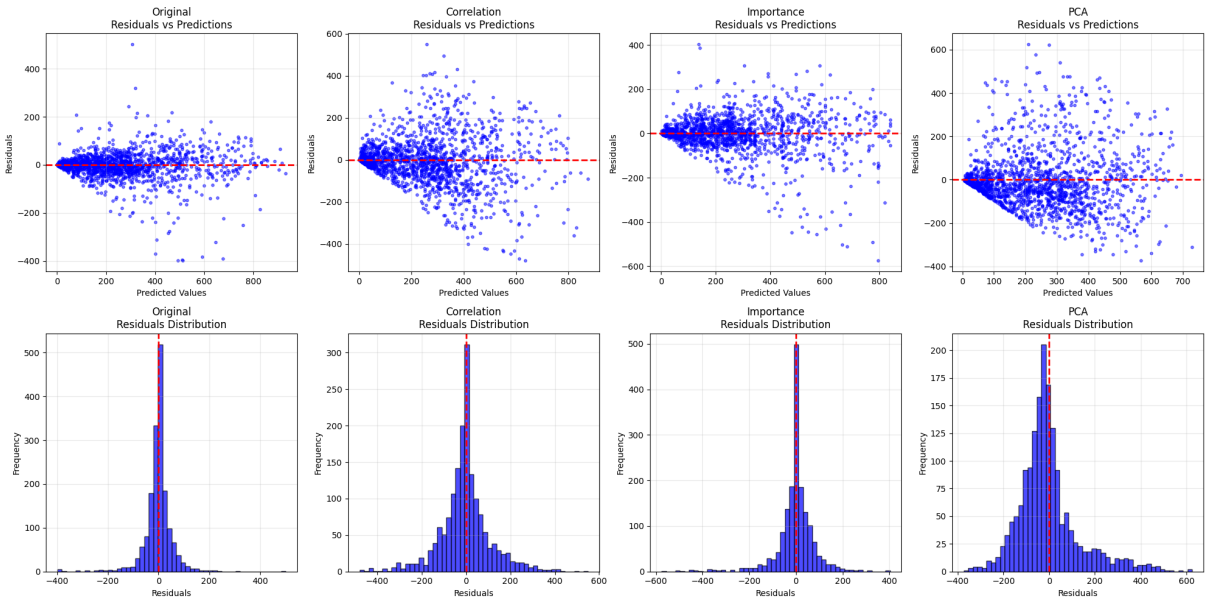
Z výsledkov vyplýva, že všetky príznaky v datasete prinášajú určitú hodnotu pre model a žiadny z nich by sa nemal odstraňovať, ak na to nie je vážny dôvod.

Porovnanie metriky R^2 na trénovacej a testovacej množine pre jednotlivé redukcie príznakov je na grafe `features_barplot`. Vedľa neho je graf, ktorý ukazuje, koľko príznakov bolo použitých v každej metóde.



Obr. 4: Porovnanie R^2 a počtu príznakov pri rôznych metódach výberu.

Graf reziduálov pre jednotlivé metódy a distribúcia reziduálov je na obrázku `features_residuals`.



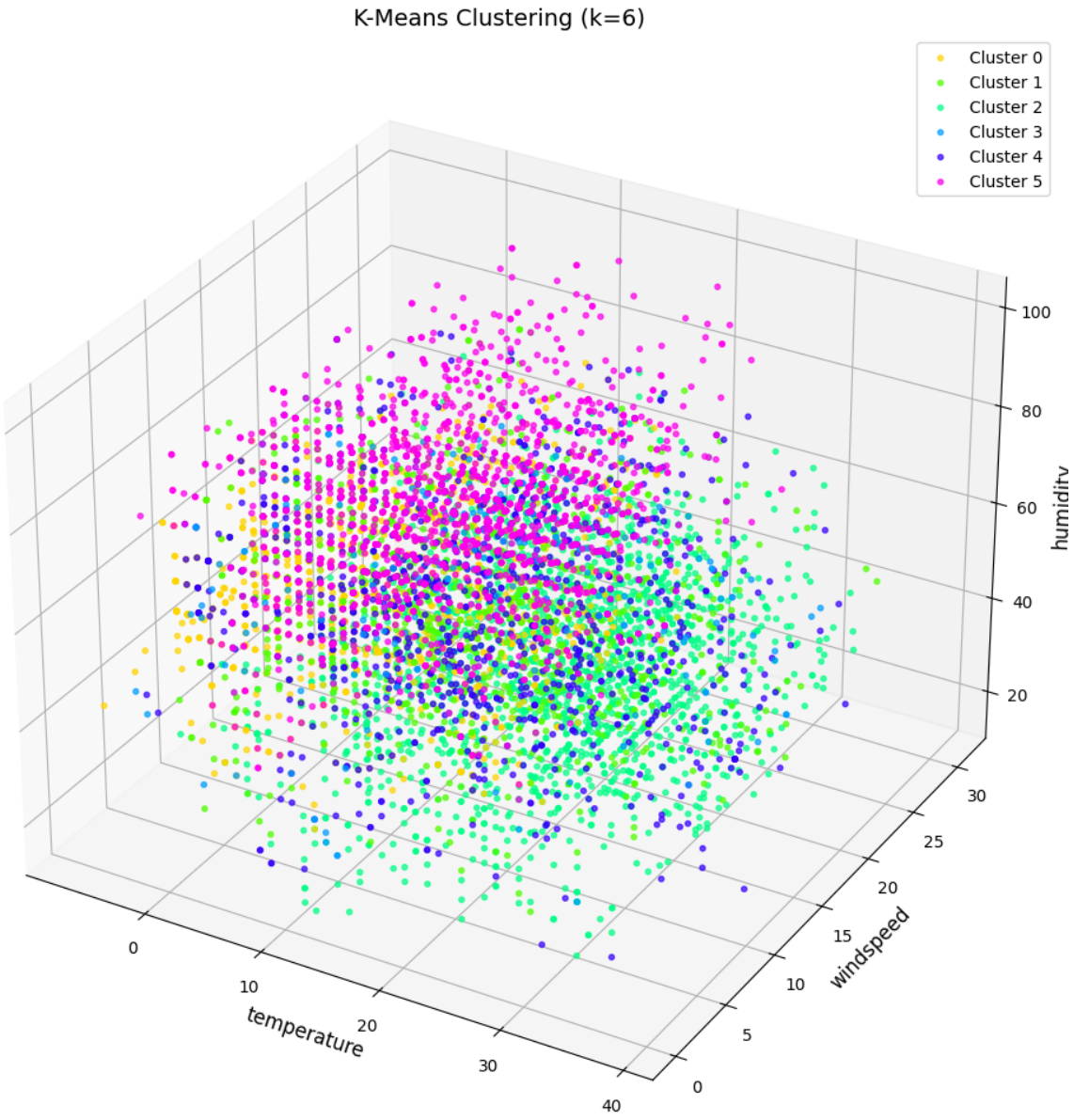
Obr. 5: Reziduály pre rôzne metódy redukcie príznakov.

Z grafu vidno, že najmenšie rozptýlenie reziduálov má model trénovaný na pôvodnej množine príznakov. Modely na zmenšených množinách majú väčšie rozptýlenie a ich reziduály sú menej symetrické okolo nuly. Výber príznakov podľa dôležitosti príznakov z

Random Forest má najmenšie rozptýlenie spomedzi zmenšených množín, no výrazne podhodnocuje vysoké hodnoty. To vidno aj v grafe distribúcie reziduálov, kde sú reziduály zošikmené zprava.

7 Zhlukovanie dát

Dáta som zhlukoval pomocou **KMeans** do 6 zhlukov. V 3D grafe sú zhluky zobrazené farbami. Na osi som dal tri spojité atribúty: `humidity`, `windspeed` a `temperature`. Graf je na obrázku `clusters`.



Obr. 6: Zhlukovanie dát pomocou KMeans.

Nevidíme jasne oddelené zhluky, čo indikuje, že dáta sú pomerne rovnomerne rozložené v priestore a nie sú prirodzene zhlukované.

Najlepší model, Random Forest, som následne natrénoval na jednotlivé zhluky a porovnal predikcie s pôvodným modelom. Tabuľka s výsledkami aj váhovaným priemerom je na obrázku `clusters_vs_original_table`.

Models Trained on Clusters vs Original Model

Model	Train Size	Test Size	Train R2	Test R2	Train RMSE	Test RMSE
Random Forest (Original)	6820	1705	0.989	0.925	22.22	57.00
Random Forest (Cluster 0)	1354	341	0.992	0.941	16.59	44.02
Random Forest (Cluster 1)	918	236	0.989	0.957	20.01	39.62
Random Forest (Cluster 2)	2054	507	0.988	0.938	21.90	49.07
Random Forest (Cluster 3)	206	54	0.971	0.834	29.50	73.93
Random Forest (Cluster 4)	892	239	0.990	0.955	20.41	45.05
Random Forest (Cluster 5)	1396	328	0.978	0.836	27.57	75.15
Random Forest (Weigh Avg)	-	1705	-	0.921	-	51.99

Obr. 7: Porovnanie výkonu modelu na jednotlivých zhlukoch.

Pre väčšinu zhlukov model dosiahol lepšie výsledky ako pôvodný model. Iba v zhlukoch 3 a 5 bol výsledok horší. To môže byť preto, že tieto zhluky obsahujú dosť málo dát, čo spôsobuje pretrénovanie.

Váhovaný priemer všetkých zhlukov dosiahol $R^2 = 0.921$, čo je síce horšie ako pôvodný model s $R^2 = 0.925$, no test RMSE sa znížilo z 57.00 na 51.99, čo je výrazné zlepšenie.

Podrobnejšia analýza jednotlivých zhlukov je v tabuľke `clusters_analysis_table`.

Cluster Characteristics Analysis

Cluster	Samples	Percentage	Avg Count	Std Count	Avg Hour	Avg Temp	Avg Hum	Weather	Test R2
0	1695	19.9%	133.8	189.0	4.6	13.2°C	68.5%	clear	0.941
1	1154	13.5%	215.8	194.6	11.2	15.1°C	59.5%	clear	0.957
2	2561	30.0%	337.7	199.2	16.5	19.2°C	47.7%	clear	0.938
3	260	3.0%	185.5	176.4	11.6	14.0°C	61.5%	clear	0.834
4	1131	13.3%	248.6	209.0	11.5	15.8°C	59.6%	clear	0.955
5	1724	20.2%	190.9	186.9	10.9	14.1°C	77.6%	cloudy	0.836

Obr. 8: Analýza jednotlivých zhlukov.

8 Neuronová sieť

Keďže ide o regresiu, výstupom neurónovej siete je jedna spojitá hodnota – predikovaný počet požičiavaní bicyklov.

Architektúra siete obsahuje:

- vstupná vrstva s 9 neurónmi (9 príznakov po spracovaní dát),
- tri skryté vrstvy so 128, 64 a 32 neurónmi,
- výstupná vrstva s 1 neurónom (predikcia spojitej hodnoty),
- aktivačná funkcia ReLU v každej skrytej vrstve.

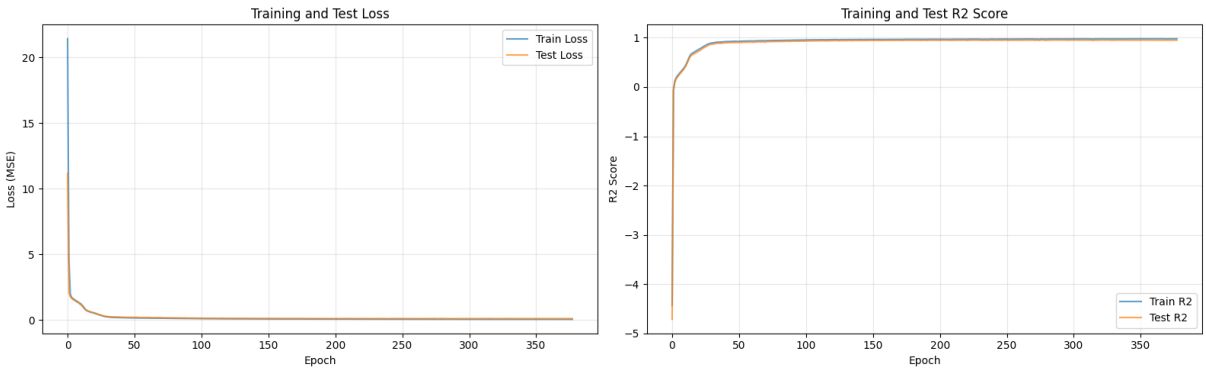
Ako optimalizátor som použil **Adam**. Stratovú funkciu som zvolil **Mean Squared Error (MSE)**, keďže sa jedná o regresiu.

Hyperparametre

Zvolené hyperparametre sú uvedené v tabuľke.

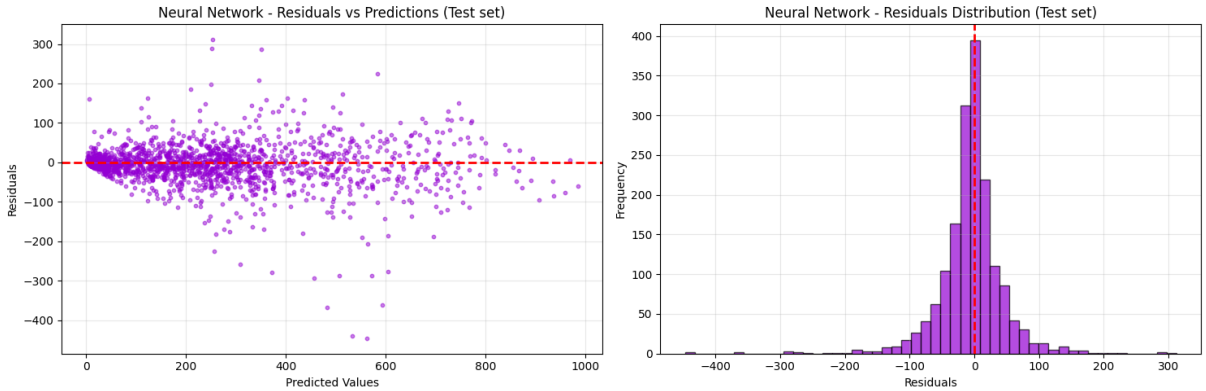
Hyperparameter	**Hodnota**
Learning Rate	1e-3
Batch Size	256
Epochs	500
Early Stopping Patience	30

Priebeh trénovania siete a vývoj R^2 skóre je na obrázku `training_curves`.



Obr. 9: Priebeh trénovania neurónovej siete.

Trénovanie sa zastavilo po 406 epochách vďaka early stoppingu, keďže sa R^2 skóre na validačnej množine prestalo zlepšovať. Model dosiahol na testovacích dátach $R^2 = 0.9379$ a RMSE $= 53.87$.



Obr. 10: Reziduály neurónovej siete.

Graf reziduálov ukazuje, že reziduály sú symetrické okolo nuly a väčšina bodov je blízko nule, čo indikuje dobrý model. Distribúcia reziduálov je približne normálna, s miernym zošikmením doprava, čo opäť indikuje, že model má tendenciu podhodnocovať veľmi vysoké hodnoty.

Zdroje a študijné materiály

Počas spracovania projektu som čerpal poznatky a inšpiráciu z nasledujúcich online videí, ktoré mi pomohli lepšie pochopiť prácu s modelmi strojového učenia, redukciu dimenzie

a analýzu dát. Tieto zdroje neboli priamo citované v texte, slúžili však ako doplnkový študijný materiál.

- StatQuest with Josh Starmer – *StatQuest: PCA main ideas in only 5 minutes!!!*, YouTube, 2018. Dostupné na: `https://www.youtube.com/watch?v=HMOI_lkzW08`
- StatQuest with Josh Starmer – *Support Vector Machines Part 1 (of 3): Main Ideas!!!*, YouTube, 2019. Dostupné na: `https://www.youtube.com/watch?v=efR1C6CvhmE`
- Visually Explained – *Support Vector Machine (SVM) in 2 minutes*, YouTube, 2021. Dostupné na: `https://www.youtube.com/watch?v=_YPScrckx28`
- Deepia – *Latent Space Visualisation: PCA, t-SNE, UMAP | Deep Learning Animated*, YouTube, 2024. Dostupné na: `https://www.youtube.com/watch?v=o_cAOa5fMhE`
- Code_Monarch – *How Principal Component Analysis (PCA) Works - AI Explained!*, YouTube Shorts. Dostupné na: `https://www.youtube.com/shorts/3YT17tXjzWI`

Vyhlásenie

Pri vypracovaní zadania som použil *ChatGPT (GPT-5, OpenAI)* na pridanie diakritiky a formátovanie niektorých častí (najme tabuľky) do formátu TeX. Obsah, interpretácie a výsledky analýz sú však mojou autorskou prácou.