Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra kybernetiky

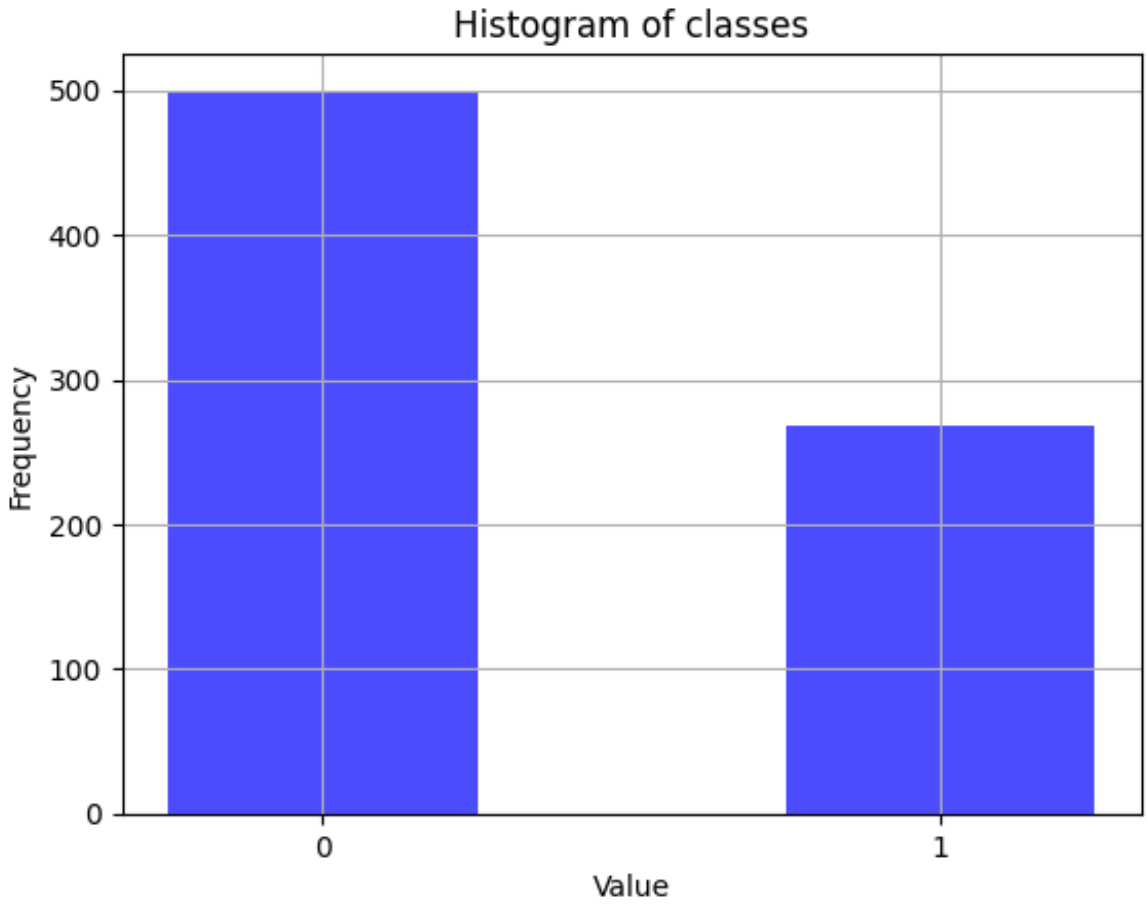
KKY/ZSY

Predikce diabetu

Yauheni Petrachenka
15. května 2024

1 Dataset

Dataset pochází s platformy kaggle. Dataset je osazen v csv souboru. Csv soubor obsahuje 9 sloupců: počet těhotností, glukóza, krevny tlak, Tloušťka kožní řasy tricepsu, inzulín, BMI, funkce rodokmenu diabetu, věk, výsledek. Celkový počet dat je roven 768. Tento soubor dat pochází původně z Národního institutu pro diabetes a onemocnění trávicího traktu a ledvin. Cílem je na základě diagnostických měření předpovědět, zda má pacient diabetes. Zkoumáním četnosti tříd byl obdržen histogram četnosti (obr. 1). Díky tomu že máme dostatečně dát každé třídy při rozdělení dat můžeme zanedbat parametrem 'stratify'.



Obrázek 1: Histogram tříd

Zdroj dat: `https://www.kaggle.com/datasets/mathchi/diabetes-data-set`

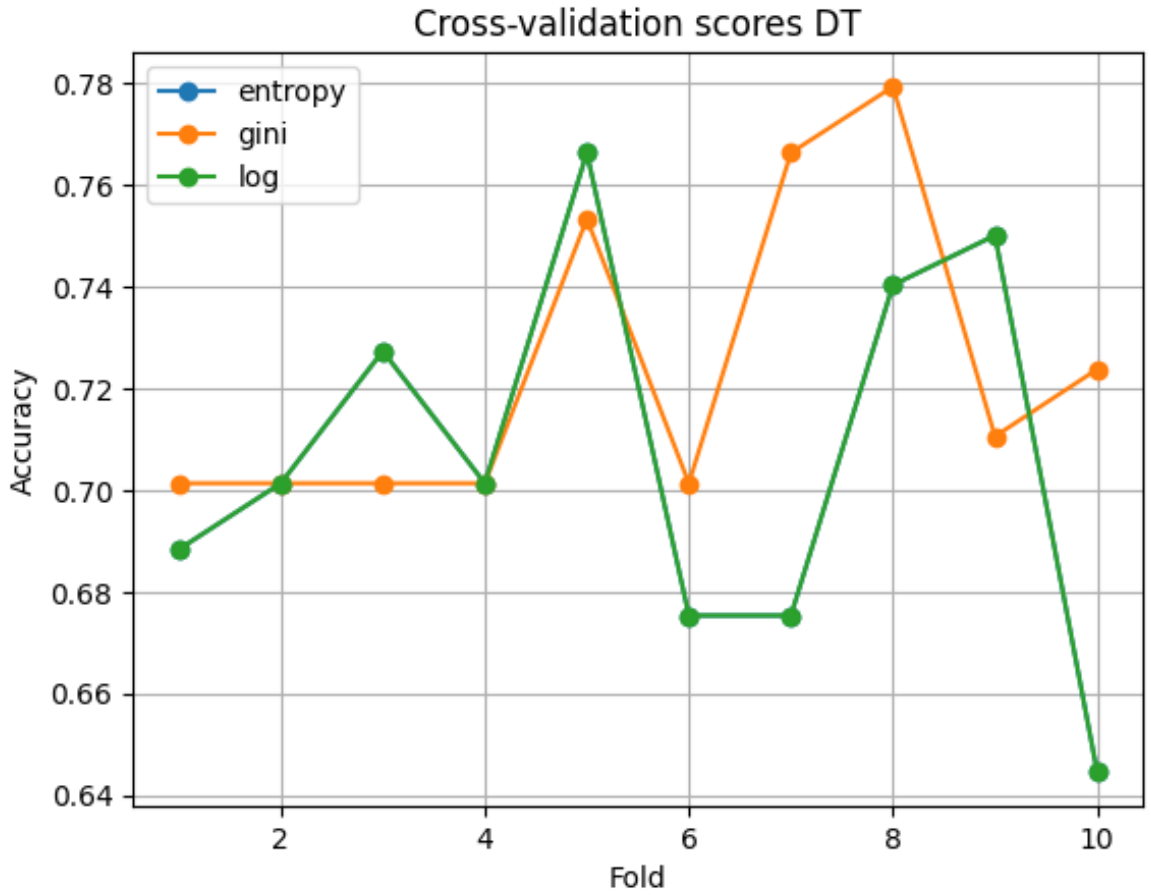
2 Rozhodovací strom

2.1 Výběr kritéria

Budeme vybírat kritérium mezi následujícími kritérii:

1. entropy $(-\sum_{i=1}^{n} p_i \cdot log_2 p_i)$
2. gini $(1 - \sum_{i=1}^{n} (P_i^2))$
3. log $(-\frac{1}{N} \sum_{i=1}^{N} y_i log(p_i) + (1 - y_i) log(1 - p_i))$

Pro zjištění nejlepšího kritéria budou vytvoření tří rozhodovacích stromů, každý strom má svůj kritérium. Abychom mohli provést spravedlivé porovnání u každého s těchto stromů inicializujeme stejný random state (V našem případě zvolíme 24). Pro učení rozhodovacích stromů zvolíme metodu cross-validation, která je realizována v knihovně sklearn KFolds. Počet vzorků při použití cross-validaion je roven 10.



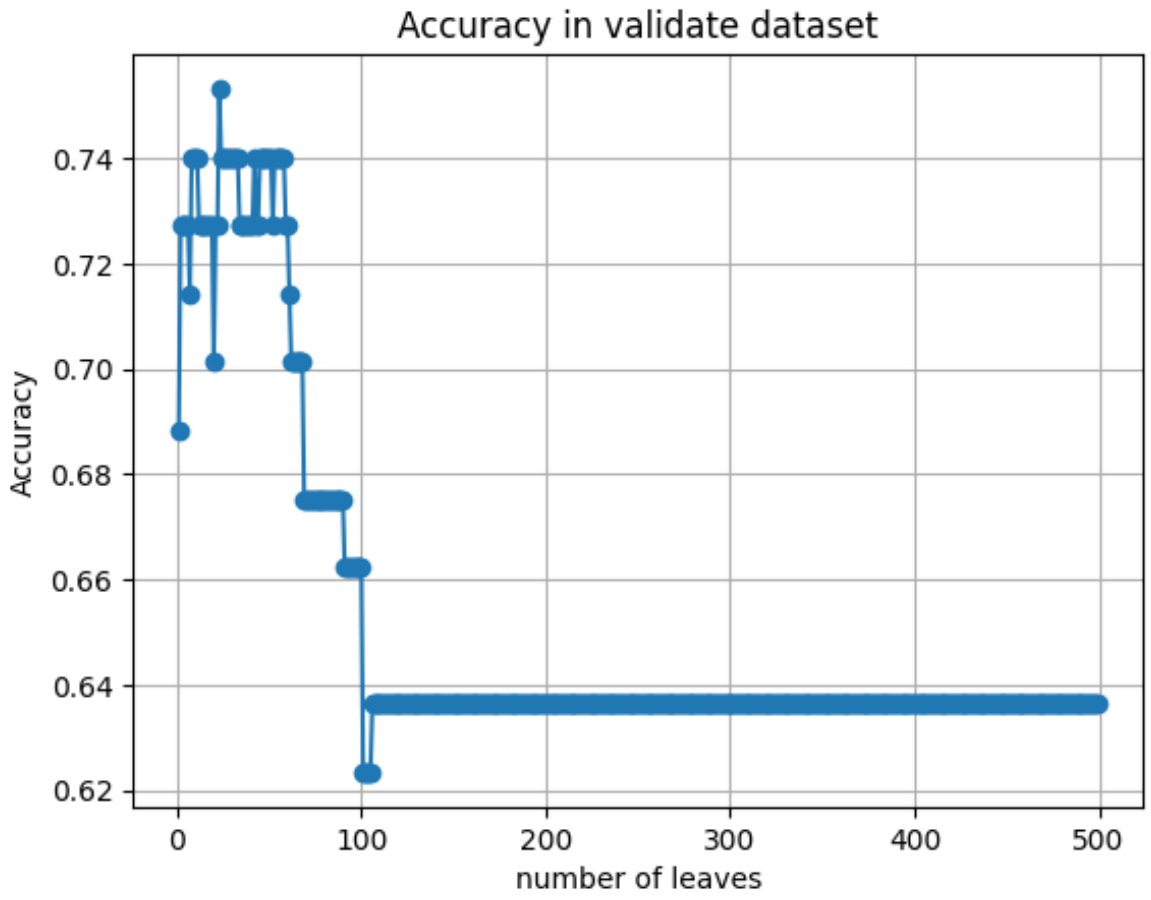
Obrázek 2: Přesnost rozhodovacích stromů

Po učení rozhodovacích stromů dostaneme obrázek (2) . Který ukazuje že přesnost pro rozhodovací strom který používal parametr 'entropy' a pro rozhodovací strom který

používal parametr 'log function' je stejná. Taky ve výsledku byly obdrženy střední přesnosti pro každý parametr: 'Mean entropy cross-validarion 0.7070061517429939', 'Cross-validation gini mean score 0.7239405331510594', 'Cross-validation log mean score 0.7070061517429939'. Ve výsledku bylo zjištěno že nejlepší kritérium je gini. Takovým způsobem byl získán parametr který bude použit na následujících modelech.

2.2 Výběr počtu listů rozhodovacího stromu

Pro analýzu daného parametru musíme předem rozdělit data na trénovací a validační část. Validační část bude obsahovat 10 procentů od všech dat. Bude zkoumán počet listů od 1 do 500. Postupní tvorbou nových rozhodovacích stromů v cyklu fór prozkoumáme všechny možnost a ve výsledku obdržíme obrázek číslo 3:



Obrázek 3: Přesnost rozhodovacích stromů

Je vidět že přesnost roste a pak výrazně klesá nejspíš to je spojeno s přeučení rozhodovacího stromu. Taky ve výsledku byly obdržení nejlepší přesnost a nejlepší počet list: 'Maximum is 0.7532467532467533', 'Best number of the nodes 24'.

2.3 Vyhodnocení modelu

Už máme potřebné parametry pro stavbu rozhodovacího stromu a vyhodnocení daného modelu. Pro dané účely budou používány funkce s balíčku sklearn: accuracy_score a classification_report.

Classification DT reort			precision	recall	f1-score	support
Healthy	0.82	0.80	0.81	50		
Diabetes	0.64	0.67	0.65	27		
accuracy			0.75	77		
macro avg	0.73	0.73	0.73	77		
weighted avg	0.76	0.75	0.75	77		

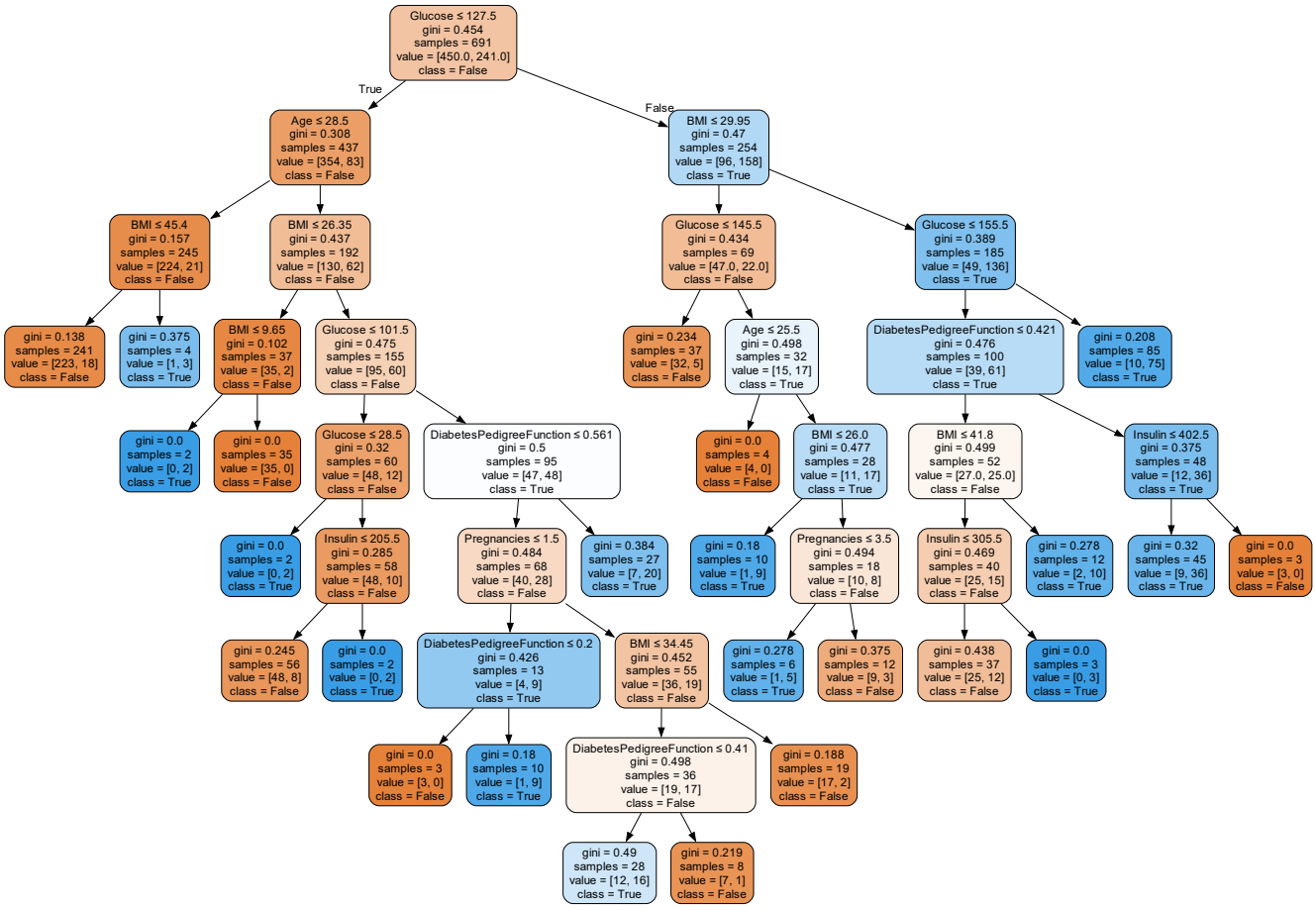
Obrázek 4: Přesnost rozhodovacího stromu

Obrázek číslo 4 nám uvádí přesnost modelu podle každé třídy.

A můžeme udělat závěr že, model lépe zvládá identifikaci zdravých lidí ve srovnání s identifikací lidí s diabetem. To může být spojeno s lepší rozlišitelností znaků zdraví ve srovnání se znaky diabetu nebo menším zastoupením případů diabetu v datech.

Přestože celková přesnost je slušná, u třídy "Diabetes"je pozorována nižší přesnost a úplnost. To může naznačovat potřebu zlepšení modelu, možná prostřednictvím inženýrství znaků, vyvážení tříd nebo použitím jiných klasifikačních algoritmů.

Pomocí knihovní graphiz můžeme vykreslit výslední rozhodovací strom:


Glucose ≤ 127.5
gini = 0.454
samples = 691
value = [450.0, 241.0]
class = False
True
False
Age ≤ 28.5
gini = 0.308
samples = 437
value = [354, 83]
class = False
BMI ≤ 29.95
gini = 0.47
samples = 254
value = [96, 158]
class = True
BMI ≤ 45.4
gini = 0.157
samples = 245
value = [224, 21]
class = False
BMI ≤ 26.35
gini = 0.437
samples = 192
value = [130, 62]
class = False
Glucose ≤ 145.5
gini = 0.434
samples = 69
value = [47.0, 22.0]
class = False
Glucose ≤ 155.5
gini = 0.389
samples = 185
value = [49, 136]
class = True
gini = 0.138
samples = 241
value = [223, 18]
class = False
gini = 0.375
samples = 4
value = [1, 3]
class = True
BMI ≤ 9.65
gini = 0.102
samples = 37
value = [35, 2]
class = False
Glucose ≤ 101.5
gini = 0.475
samples = 155
value = [95, 60]
class = False
gini = 0.234
samples = 37
value = [32, 5]
class = False
Age ≤ 25.5
gini = 0.498
samples = 32
value = [15, 17]
class = True
DiabetesPedigreeFunction ≤ 0.421
gini = 0.476
samples = 100
value = [39, 61]
class = True
gini = 0.208
samples = 85
value = [10, 75]
class = True
gini = 0.0
samples = 2
value = [0, 2]
class = True
gini = 0.0
samples = 35
value = [35, 0]
class = False
Glucose ≤ 28.5
gini = 0.32
samples = 60
value = [48, 12]
class = False
DiabetesPedigreeFunction ≤ 0.561
gini = 0.5
samples = 95
value = [47, 48]
class = True
gini = 0.0
samples = 4
value = [4, 0]
class = False
BMI ≤ 26.0
gini = 0.477
samples = 28
value = [11, 17]
class = True
BMI ≤ 41.8
gini = 0.499
samples = 52
value = [27.0, 25.0]
class = False
Insulin ≤ 402.5
gini = 0.375
samples = 48
value = [12, 36]
class = True
gini = 0.0
samples = 2
value = [0, 2]
class = True
Insulin ≤ 205.5
gini = 0.285
samples = 58
value = [48, 10]
class = False
Pregnancies ≤ 1.5
gini = 0.484
samples = 68
value = [40, 28]
class = False
gini = 0.384
samples = 27
value = [7, 20]
class = True
gini = 0.18
samples = 10
value = [1, 9]
class = True
Pregnancies ≤ 3.5
gini = 0.494
samples = 18
value = [10, 8]
class = False
Insulin ≤ 305.5
gini = 0.469
samples = 40
value = [25, 15]
class = False
gini = 0.278
samples = 12
value = [2, 10]
class = True
gini = 0.32
samples = 45
value = [9, 36]
class = True
gini = 0.0
samples = 3
value = [3, 0]
class = False
gini = 0.245
samples = 56
value = [48, 8]
class = False
gini = 0.0
samples = 2
value = [0, 2]
class = True
DiabetesPedigreeFunction ≤ 0.2
gini = 0.426
samples = 13
value = [4, 9]
class = True
BMI ≤ 34.45
gini = 0.452
samples = 55
value = [36, 19]
class = False
gini = 0.278
samples = 6
value = [1, 5]
class = True
gini = 0.375
samples = 12
value = [9, 3]
class = False
gini = 0.438
samples = 37
value = [25, 12]
class = False
gini = 0.0
samples = 3
value = [0, 3]
class = True
gini = 0.0
samples = 3
value = [3, 0]
class = False
gini = 0.18
samples = 10
value = [1, 9]
class = True
DiabetesPedigreeFunction ≤ 0.41
gini = 0.498
samples = 36
value = [19, 17]
class = False
gini = 0.188
samples = 19
value = [17, 2]
class = False
gini = 0.49
samples = 28
value = [12, 16]
class = True
gini = 0.219
samples = 8
value = [7, 1]
class = False

3 Náhodný les

3.1 Výběr nejlepších parametrů

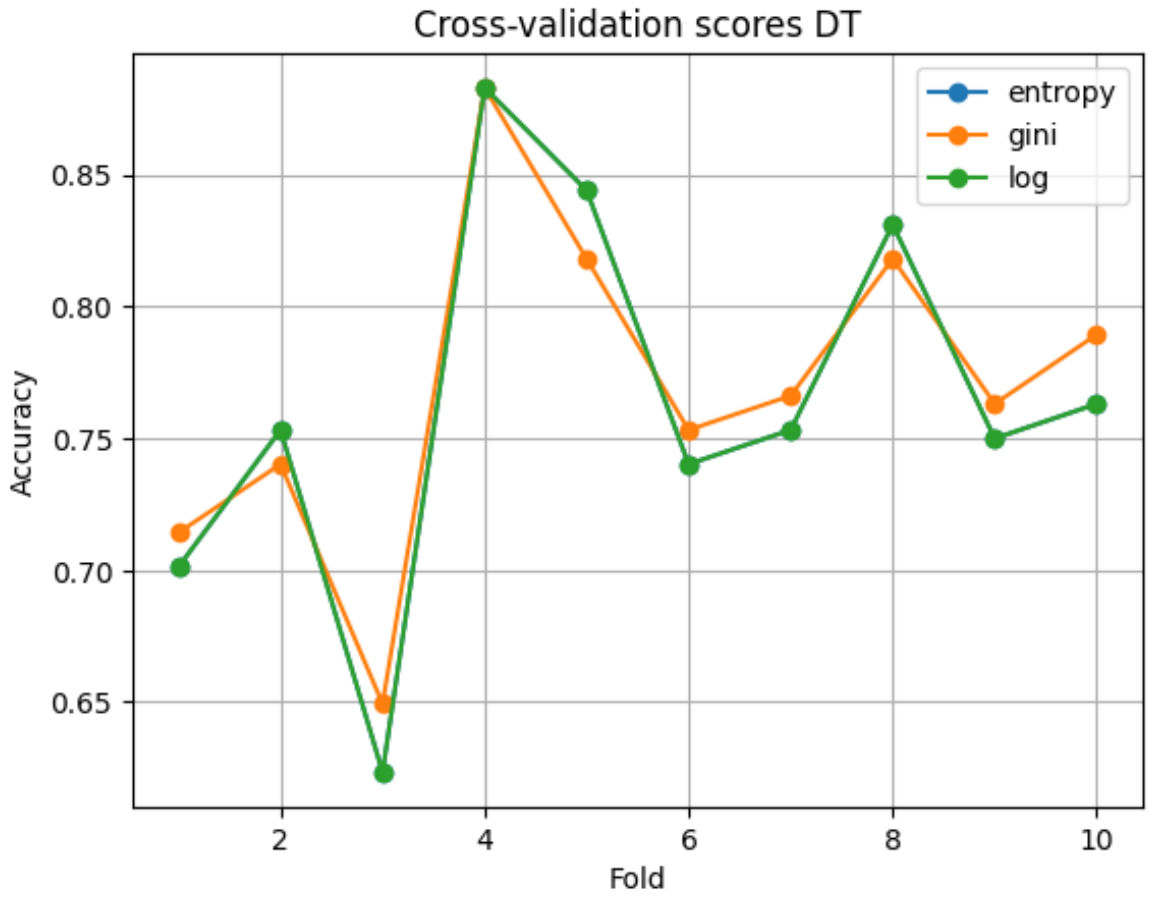
Pro výběr rozhodovacích parametrů v náhodném lese byla použítá funkce GridSearchCV() . Množina parametrů mezi kterými byla vyhledávána optimální kombinace:

1. 'n_estimators': [50, 100, 200],
2. 'max_depth': [5, 10, 20],
3. 'min_samples_split': [2, 5, 10],
4. 'min_samples_leaf': [1, 2, 4],
5. 'max_features': ['sqrt', 'log2']

Takovým způsobem byly zjištěni následující nejlepší parametry pro náhodný les:

1. max_depth=20
2. max_features=sqrt
3. min_samples_leaf=1
4. min_samples_split=10
5. n_estimators=100

Taky budeme zkoumat nejlepší kritérium podobně jako pro rozhodovací strom. Nadefinujeme tří náhodných lesů a budeme je zkoumat pomocí podobného způsobu, který byl použit při výběru kritéria u rozhodovacího stromu.



Obrázek 5: Přesnost náhodného lesa

S grafu je vidět že situace s náhodným lesem je stejná, jako s rozhodovacím stromem, v našem případě nejlepší kritérium je gini, a entopie a log mají stejnou hodnotu. Obdržíme následující výpis programu: 'Mean cross-validation entropy score 0.7643028024606972', 'Mean cross-validation gini score 0.7695488721804512', 'Mean cross-validation log score 0.7643028024606972'. Pro vyhodnocení modelů budeme používat kritérium gini.

3.2 Vyhodnocení modelu

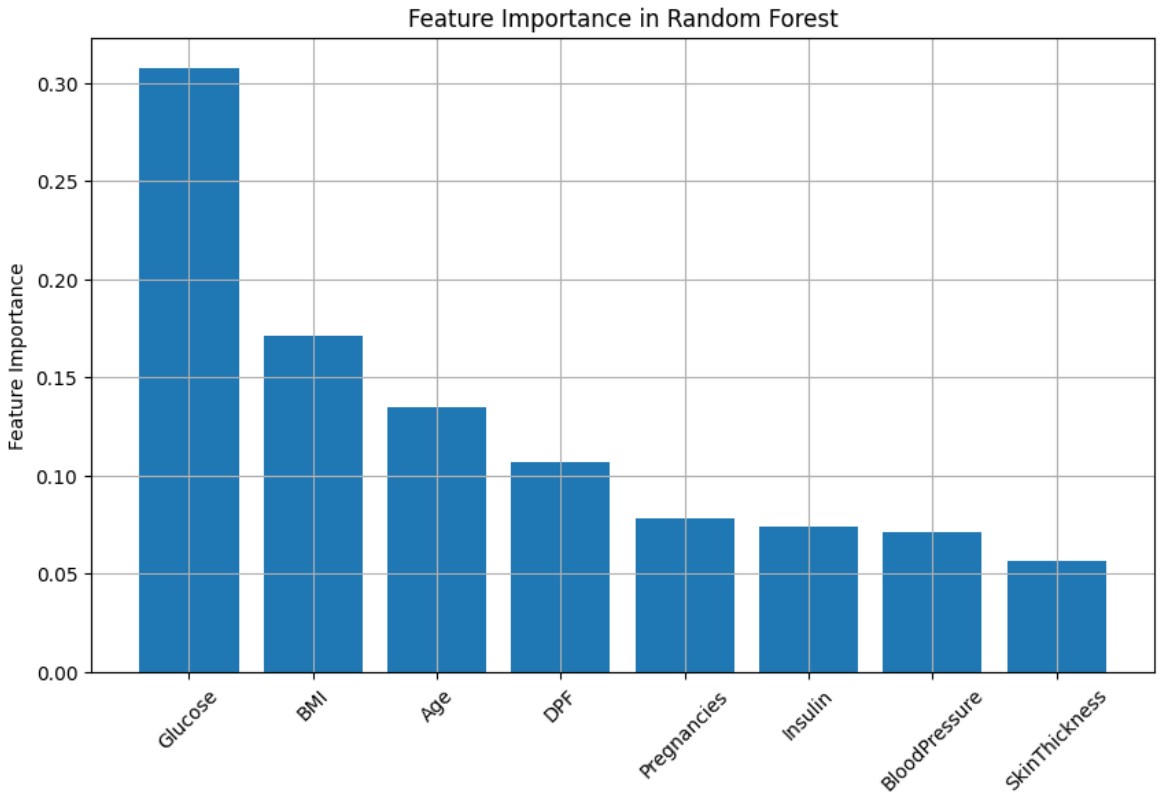
Už máme potřebné parametry pro stavbu náhodného lesu a vyhodnocení daného modelu. Pro dané účely budou používány funkce s balíčku sklearn: accuracy_score , classification_report a feature_importance_.

Classification RF reort:

	precision	recall	f1-score	support
Healthy	0.84	0.82	0.83	50
Diabetes	0.68	0.70	0.69	27
accuracy			0.78	77
macro avg	0.76	0.76	0.76	77
weighted avg	0.78	0.78	0.78	77

Obrázek 6: Přesnost náhodného lesa

Model náhodného lesa ukázal dobré výsledky, zejména pro třídu "Healthy", kde jsou ukazatele přesnosti, úplnosti a F1 skóre vyšší než pro třídu "Diabetes". Ukazatele pro třídu "Diabetes"jsou nižší, což může svědčit o složitosti klasifikace tohoto stavu nebo o potřebě dalšího nastavení modelu pro zlepšení jeho výkonu pro tuto třídu. Celková přesnost modelu, makro a vážené průměry jsou dostatečně vysoké, což svědčí o dobrém celkovém výkonu modelu na prezentovaných datech.



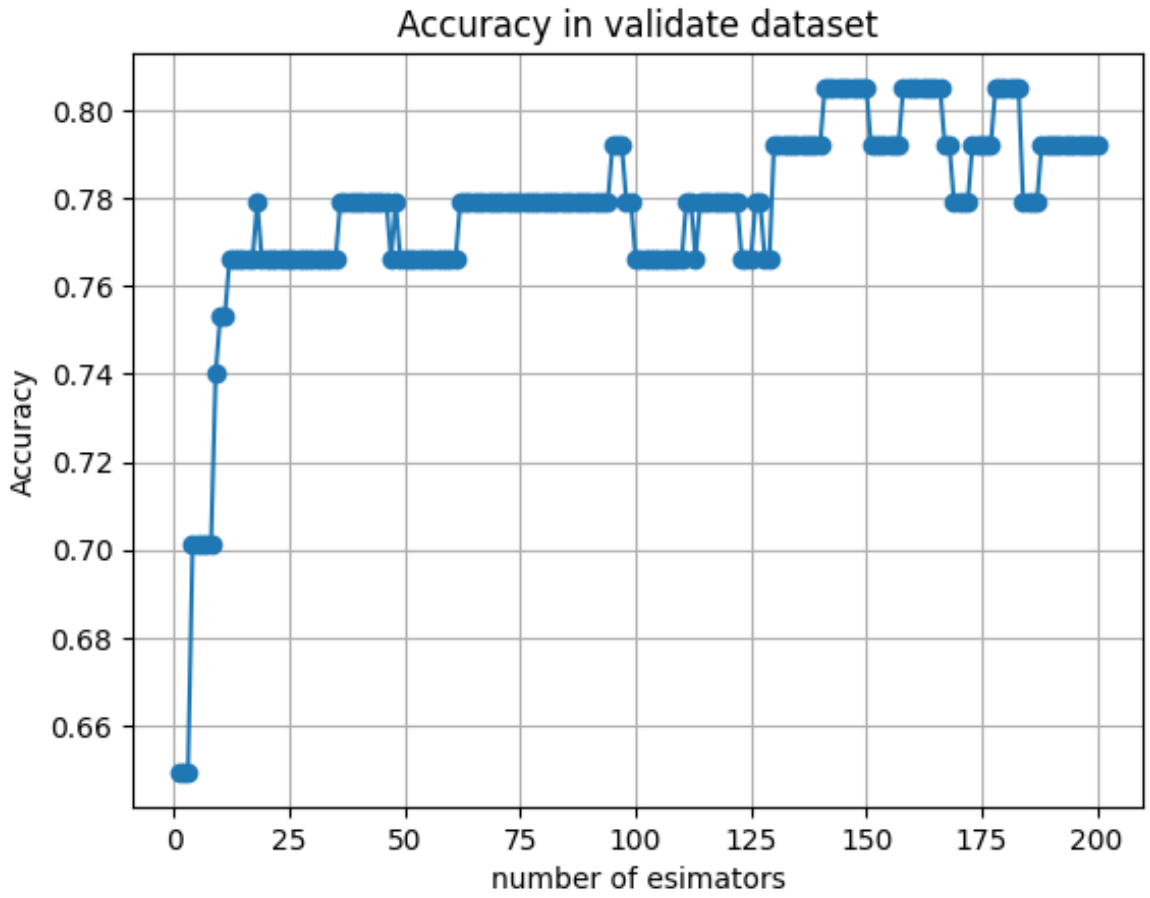
Obrázek 7: Feature importance

Na předloženém diagramu (obr. 7) je znázorněn význam rysů v modelu náhodného lesa, který se používá pro predikci diabetu. Model se silně spoléhá na úroveň glukózy, což potvrzuje jeho význam pro diagnostiku diabetu. BMI a věk jsou také významnými faktory, což odpovídá jejich známému vlivu na zdraví člověka a riziko rozvoje diabetu. Méně významné faktory, jako je hladina inzulínu a počet těhotenství, mohou vyžadovat další analýzu pro hodnocení jejich role v kontextu konkrétních klinických údajů.

4 Gradient Boosting

4.1 Výběr nejlepších parametrů

Pro výběr nejlepšího počtu rozhodovacích stromů postup bude podobný jako při výběru nejlepšího počtu listů u rozhodovacích stromů. Rozsah počtu rozhodovacích stromů bude se rovnat od 1 do 200. Po provedení všech iterací ve výsledku dostaneme graf 8.



Obrázek 8: Počet rozhodovacích stromů

S grafu je vidět že přesnost na začátku roste a pak začíná kmitat. Taky byly obdrženy údaje o nejlepším počtu stromů a o nejlepší přesnosti: 'The best accuracy is 0.8051948051948052', 'The best number of estimators 140'.

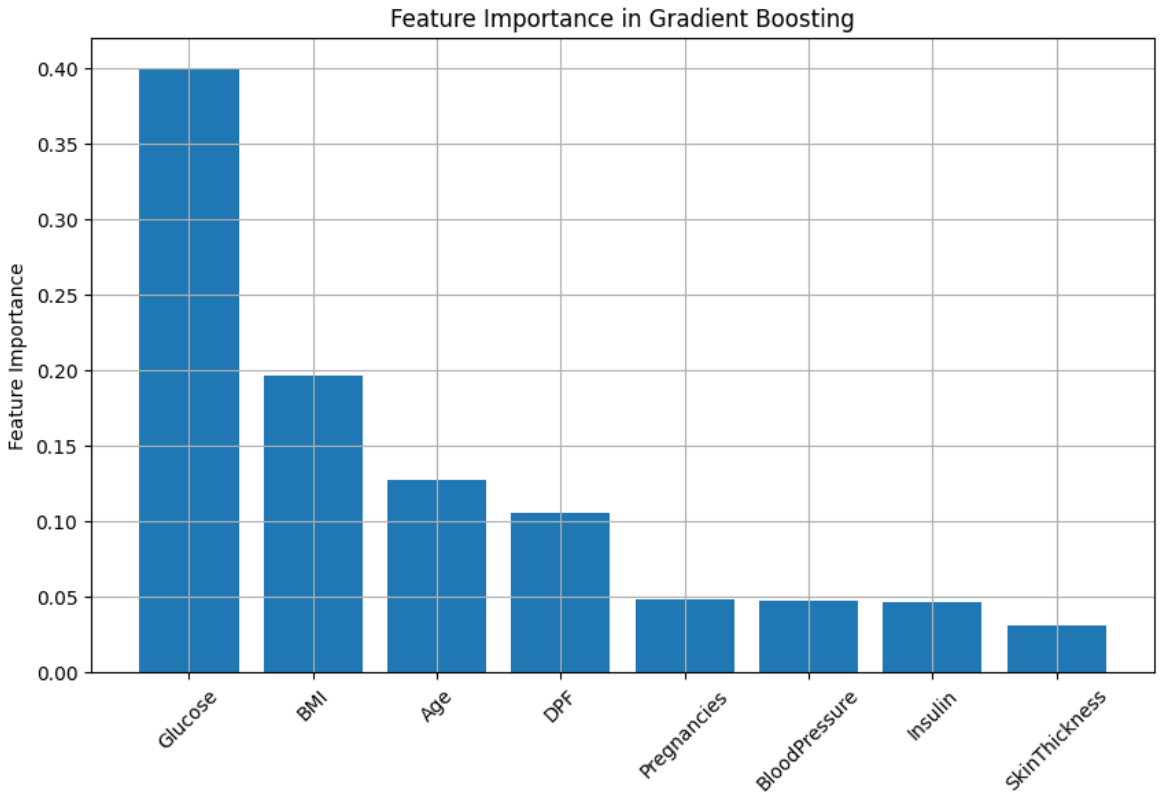
4.2 Vyhodnocení modelu

Classification GB reort:

	precision	recall	f1-score	support
Healthy	0.87	0.80	0.83	50
Diabetes	0.68	0.78	0.72	27
accuracy			0.79	77
macro avg	0.77	0.79	0.78	77
weighted avg	0.80	0.79	0.80	77

Obrázek 9: Přesnost modelu

Model gradientního zvýšení ukázal dobré výsledky, zejména pro třídu "Healthy", kde ukazatele přesnosti, úplnosti a F1 skóre jsou vyšší než pro třídu "Diabetes". Ukazatele pro třídu "Diabetes"jsou nižší, což může svědčit o složitosti klasifikace tohoto stavu nebo o potřebě dalšího nastavení modelu pro zlepšení jeho výkonu pro tuto třídu. Celková přesnost modelu a vážené průměry jsou dostatečně vysoké, což svědčí o dobrém celkovém výkonu modelu na prezentovaných datech.



Obrázek 10: Feature importance

Model velmi oceňuje faktory, jako je hladina glukózy, BMI a věk, což odpovídá jejich známému vlivu na riziko diabetu. Rysy související s celkovým stavem zdraví a metabolickými procesy, jako je BMI a věk, také významně ovlivňují předpovědi modelu. Méně významné rysy mohou vyžadovat další analýzu pro určení jejich role v konkrétních podmínkách nebo nastaveních modelu.

5 Závěr

Semestrální práce zaměřená na predikci diabetu pomocí různých modelů strojového učení ukázala několik klíčových poznatků a výsledků. Analýza rozhodovacích stromů, náhodných lesů a gradientního zvýšení na datasetu z platformy Kaggle poskytla podrobný pohled na výkon jednotlivých modelů s různými parametry. Bylo zjištěno, že kritérium Gini přineslo nejlepší výsledky pro rozhodovací stromy i náhodné lesy, což naznačuje jeho efektivitu v rozhodovacích algoritmech v rámci tohoto konkrétního datasetu. Výběr optimálního počtu listů a hloubky stromů byl důležitý pro zajištění dostatečné přesnosti bez rizika přeučení modelu.

Gradientní zvýšení, stejně jako ostatní testované modely, ukázalo lepší schopnosti při identifikaci zdravých jedinců ve srovnání s diagnostikou diabetu, což může signalizovat potřebu dalšího vývoje modelu pro zlepšení diagnostiky diabetu. Obecně modely

prokázaly solidní celkovou přesnost, což bylo doplněno analýzou významnosti rysů, kde hladina glukózy, BMI a věk byly identifikovány jako klíčové prediktivní faktory.

Závěrem, tato semestrální práce přispívá k lepšímu pochopení možností a omezení aktuálních metod strojového učení v predikci diabetu a zdůrazňuje důležitost pečlivého výběru modelů a jejich parametrů pro zlepšení diagnostických schopností v medicínských aplikacích.