ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl

WYDZIAŁ
ELEKTROTECHNIKI
I INFORMATYKI
POLITECHNIKI RZESZOWSKIEJ

Projekt

z przedmiotu

Język Python w analizie danych

Temat projektu: Porównanie metod imputacji oraz algorytmów uczenia maszynowego na podstawie klasyfikacji

Prowadzący: mgr inż. Patryk Organiściak

Imię i nazwisko: inż. Kamil Hansel

Nr albumu: 166728

Rok studiów: 1 FS0-DU

Grupa projektowa: P2

Data oddania projektu: 20.01.2025

1. Opis zbioru danych

Dane wykorzystane w ramach tego projektu zostały pobrane z repozytorium Uniwersytetu Kalifornijskiego w Irvine. Zbiór dotyczy klasyfikacji grzybów decydując czy dany grzyb jest jadalny czy trujący. Zbiór danych zawiera 61068 wierszy oraz 20 kolumn (21 z cechą docelową) Parametry cech wyglądają następująco:

	name	role	type	...	description	units	missing_values
0	class	Target	Categorical	...	None	None	no
1	cap-diameter	Feature	Continuous	...	None	None	no
2	cap-shape	Feature	Categorical	...	None	None	no
3	cap-surface	Feature	Categorical	...	None	None	yes
4	cap-color	Feature	Categorical	...	None	None	no
5	does-bruise-or-bleed	Feature	Categorical	...	None	None	no
6	gill-attachment	Feature	Categorical	...	None	None	yes
7	gill-spacing	Feature	Categorical	...	None	None	yes
8	gill-color	Feature	Categorical	...	None	None	no
9	stem-height	Feature	Continuous	...	None	None	no
10	stem-width	Feature	Continuous	...	None	None	no
11	stem-root	Feature	Categorical	...	None	None	yes
12	stem-surface	Feature	Categorical	...	None	None	yes
13	stem-color	Feature	Categorical	...	None	None	no
14	veil-type	Feature	Categorical	...	None	None	yes
15	veil-color	Feature	Categorical	...	None	None	yes
16	has-ring	Feature	Categorical	...	None	None	no
17	ring-type	Feature	Categorical	...	None	None	yes
18	spore-print-color	Feature	Categorical	...	None	None	yes
19	habitat	Feature	Categorical	...	None	None	no
20	season	Feature	Categorical	...	None	None	no

Rysunek 1 - Informacje dotyczące cech: nazwa, rola, typ oraz indeks pustych wartości

Link do dostępu: https://archive.ics.uci.edu/dataset/848/secondary+mushroom+dataset

2. Transformacja danych

Dane wstępnie zostały przemapowane za pomocą funkcji preprocessing.OrdinalEncoder() z pakietu sklearn. Aby zbiór mógł zostać poddany procesowi imputacji zmieniono typ kolumn zbioru na float. Następnie obliczono współczynniki pustych wartości w każdej kolumnie, tak aby można było wykluczyć atrybuty mało istotne. Wprowadzony próg dopuszczalności ustalono na 50%. Następnie wyrzucono kolumny, w których ponad połowa danych zawierała puste wartości oraz statycznie wprowadzono wektor_pominiec który zostanie wykorzystany w funkcji wspomagającej jeden algorytm imputacji. Następnie podzielono zbiór danych na zbiór treningowy oraz zbiór testowy.

3. Imputacja danych

W projekcie wykorzystano 4 różne sposoby imputacji danych:

- Imputacja zastępująca puste wartości modą(dominantą) danej kolumny,
- imputacja zastępująca puste wartości medianą danej kolumny,
- imputacja za pomocą algorytmu IterativeImputer za pomocą którego wartości wprowadzone zostały następnie zaokrąglone do liczb całkowitych w kolumnach, które zawierają dane kategoryczne,
- imputacja za pomocą algorytmu KNNImputer.

Ponadto zaprogramowano opcję umożliwiającą pominięcie przeprowadzenia imputacji ze względu na fakt, że dwa z pięciu algorytmów uczenia maszynowego wykorzystanego w tym projekcie jest w stanie obsłużyć braki danych w kolumnach. Wszystkie sposoby imputacji zostały opatrzone w funkcje mierzące czas potrzebny na przetworzenie kodu.

4. Dopasowanie modeli uczenia maszynowego oraz przeprowadzanie predykcji

W projekcie wykorzystano 5 różnych algorytmów uczenia maszynowego:

- Drzewo decyzyjne
- Regresja logistyczna
- Naiwny klasyfikator bayesowski
- SVM
- HistGradientBoostingClassifier

Podobnie jak przy imputacji, wszystkie algorytmy uczenia maszynowego zostały opatrzone w funkcje mierzące czas potrzebny na przetworzenie kodu. Ponadto obliczono stopień dopasowania każdego modelu. Na koniec zebrano wszystkie wyniki i przedstawiono je w tabeli poniżej.

5. Podsumowanie wyników

Z podanych wyników można wywnioskować, że metody imputacji danych nie wpływają w znacznym stopniu na poziom predykcji, natomiast wpływają one w znaczniejszym stopniu na czas obliczeniowy. Algorytm KNNImputer pomimo parametru neighbours ustawionego na 1 potrzebował aż 100 sekund na przeprowadzenie imputacji. Ciekawym spostrzeżeniem jest fakt, iż największy stopień dopasowania występuje przy dwóch algorytmach które są w stanie obsłużyć braki wartości. Tabela została posortowana wobec parametru ‘Accuracy’.

	Imputation Method	Classifier	Accuracy	Total Time (s)
2	Moda	Bayes naiwny	0.591452	0.040
7	Mediana	Bayes naiwny	0.594727	0.078
12	IterativeImputer	Bayes naiwny	0.598821	1.489
17	KNNImputer	Bayes naiwny	0.599149	100.450
13	IterativeImputer	SVM	0.618361	1.812
11	IterativeImputer	Regresja logistyczna	0.618361	1.809
16	KNNImputer	Regresja logistyczna	0.625130	100.727
18	KNNImputer	SVM	0.625239	100.714
6	Mediana	Regresja logistyczna	0.626385	0.329
8	Mediana	SVM	0.626440	0.317
3	Moda	SVM	0.626549	0.280
1	Moda	Regresja logistyczna	0.627204	0.330
0	Moda	Drzewa decyzyjne	0.920910	0.157
10	IterativeImputer	Drzewa decyzyjne	0.922002	1.649
5	Mediana	Drzewa decyzyjne	0.926805	0.200
15	KNNImputer	Drzewa decyzyjne	0.948365	100.572
20	Brak	Drzewa decyzyjne	0.961083	0.179
19	KNNImputer	HistGradientBoostingClassifier	0.996016	100.974
4	Moda	HistGradientBoostingClassifier	0.997380	0.678
9	Mediana	HistGradientBoostingClassifier	0.997380	0.628
24	Brak	HistGradientBoostingClassifier	0.998472	0.585
14	IterativeImputer	HistGradientBoostingClassifier	0.998635	2.016
21	Brak	Regresja logistyczna	NaN	NaN
22	Brak	Bayes naiwny	NaN	NaN
23	Brak	SVM	NaN	NaN

Rysunek 2 - Tabela podsumowująca wyniki imputacji oraz uczenia maszynowego