UNIWERSYTET WSB MERITO WROCŁAW	
Laboratorium z Podstaw Uczenia Maszynowego	
Ćwiczenie nr	DATA:
Temat:	Eksploracyjna analiza danych
Nazwisko i imię	Kacper Studenny
Nr indeksu	81573
Grupa:	2
Prowadzący	*Prof. dr hab. Urszula Markowska-Kaczmar*

Eksploracyjna analiza danych dla zbioru…

Zbiór danych dotyczy zestawu badań nowotworów, gdzie określono końcową diagnozę jako ‘Niegroźny’ lub “Złośliwy’.

1 Opisać klasy, liczbę próbek w zbiorze i liczbę próbek w każdej klasie

```
#zliczenia klas
iloscZlosliwych=0
iloscNiegrozny=0
# klasa 4 to oznaczenie złośliwych przypadków, klasa 2 to oznaczenie zdrowych przypadków
for dana in dane['Class']:
    if(classCheck(dana,4)):
        iloscZlosliwych+=1
    else:
        iloscNiegrozny+=1

print("Ilosc Zlosliwych: "+str(iloscZlosliwych))
print("Ilość Niegroźny: " + str(iloscNiegrozny))
print("Łacznie: " + str(iloscNiegrozny+iloscZlosliwych))
```

```
Ilosc Zlosliwych: 241
Ilość Niegroźny: 457
Łacznie: 698
```

Dokonujemy zliczenia ilości klas przy pomocy pętli, która sprawdza wartość kolumny “Class” i dzięki niej określa przynależność do klasy.

Zbiór posiada dwie klasy: benign nazwany Niegroźnym (mającą 457 próbek) i malignant nazwany Złośliwym (mającą 241 próbek). W zbiorze danych klasa Niegroźny reprezentowana jest liczbą 2, a złośliwy liczbą 4.

Różnica w ilości próbek może znacząco wpłynąć na wynik uczenia. W przypadku, gdy jedna klasa ma znacznie więcej próbek niż druga, model może mieć trudności w poprawnym nauczeniu się klasyfikacji mniejszej klasy, ponieważ będzie bardziej skłonny do przewidywania większej klasy ze względu na większą ilość dostępnych przykładów.

Rozwiązać ten problem można dokonać oversampling mniejszej klasy (powielenie przykładów tej klasy) lub undersampling większej klasy (zmniejszenie liczby przykładów tej klasy). Inną techniką jest zastosowanie algorytmów wagujących przykłady, które dają większą wagę mniejszej klasie.

Można jeszcze użyć innej metody równoważenia danych zwaną undersamplingiem, który polega na zmniejszaniu liczby próbek dla bardziej licznych klas. Można to zrobić przez losowe lub celowe usunięcie pewnej liczby

próbek z tych klas. Jednak trzeba być ostrożnym, aby nie utracić istotnych informacji.

2 Sprawdzić czy występują brakujące wartości w zmiennych

```
#sprawdza czy w danej tablicy brakuje danych
def isAnyNull(tablica):
    for zmienna in tablica:
        if(zmienna=="?"):
            return True
    return False
```

```
for label in columny:
    if(isAnyNull(dane[label])):
        print("Brakuje danych w kolumnie: " + str(label))

```

```
Brakuje danych w kolumnie: Bare Nuclei
```

Sprawdziliśmy czy w jakiś kolumnach brakuje danych przy pomocy powyższego kodu. Sprawdza on kolejne kolumny w poszukiwaniu znaku “?”, który według dokumentacji oznacza brak danych. Wykryto brakujące dane tylko w kolumnie “Bare Nuclei”, co zgadza się również z dokumentacją.

Głównym problemem w naszym zestawie jest nierównomierne rozłożenie danych, co może prowadzić do poważnych konsekwencji, np. skrzywienia modelu. Jeśli jedna z klas jest znacznie bardziej liczna niż pozostałe, model może wykazywać skrzywienie i skłaniać się do przewidywania bardziej licznej klasy, ignorując mniej liczne klasy.może doprowadzić do niskiej dokładności: Jeśli modele są oceniane za pomocą prostej dokładności, może się wydawać, że model jest skuteczny, jeśli dobrze radzi sobie z przewidywaniem dominującej klasy, podczas gdy mniej liczne klasy są niedokładnie przewidywane.

Trudności w wykrywaniu rzadkich zdarzeń: Jeśli rzadkie zdarzenia są nierównomiernie rozłożone, model może mieć trudności w ich wykrywaniu i przewidywaniu.

Dane należy skorygować usuwając wybrakowane wersy, aby utrzymać pełność danych do przyszłej poprawnej nauki modelu.

3 Podsumowanie dla trzech wybranych zmiennych numerycznych

```
#Wybieramy 3 kolumny na ktorych przprowadzimy dalsze zadanie
wybaneLabel = ['Clump thickness', 'uniformity of cell size','Uniformity of Cell Shape']
```

```
#analiza danych
for z in wybaneLabel:
    print("Dane dla " + str(z))
    #średnie
    print("średnia: " + str(dane[z].mean()))
    #mediana
    print("Mediana: " + str(dane[z].median()))
    #max
    print("Maksymalna wartość wystepująca: " + str(pd.DataFrame.max(dane[z])))
    #min
    print("Minimalna wartość wystepująca: " + str(pd.DataFrame.min(dane[z])))
    #Standardowe odchylenie
    print("Standardowe odchylenie: " + str(pd.DataFrame.std(dane[z])))
    #kwartyle
    print("Kwartyle: ")
    print(str(dane[z].quantile([0.25, 0.75], interpolation='nearest')))

    print()
```

(średnia, odchylenie, mediana, wart. maks. wart. minimalna, kwartyle) .

```
Dane dla Clump thickness
średnia: 4.416905444126074
Mediana: 4.0
Maksymalna wartość wystepująca: 10
Minimalna wartość wystepująca: 1
Standardowe odchylenie: 2.8176733983653017
Kwartyle:
0.25    2
0.75    6


Dane dla uniformity of cell size
średnia: 3.1375358166189113
Mediana: 1.0
Maksymalna wartość wystepująca: 10
Minimalna wartość wystepująca: 1
Standardowe odchylenie: 3.052575308453491
Kwartyle:
0.25    1
0.75    5


Dane dla Uniformity of Cell Shape
średnia: 3.2106017191977076
Mediana: 1.0
Maksymalna wartość wystepująca: 10
Minimalna wartość wystepująca: 1
Standardowe odchylenie: 2.9728666675080633
Kwartyle:
0.25    1
0.75    5
```

Wyliczono żądane wartości z pomocą biblioteki Pandas.

Wybraliśmy do analizy 3 pierwsze wartości z tabeli, czyli ‘Clump thickness'(1), 'uniformity of cell size'(2), 'Uniformity of Cell Shape(3)’ (Grubość grudek”, „Jednorodność wielkości komórek”, „Jednorodność kształtu komórek”). Dla tych trzech parametrów przeprowadziliśmy analizę.

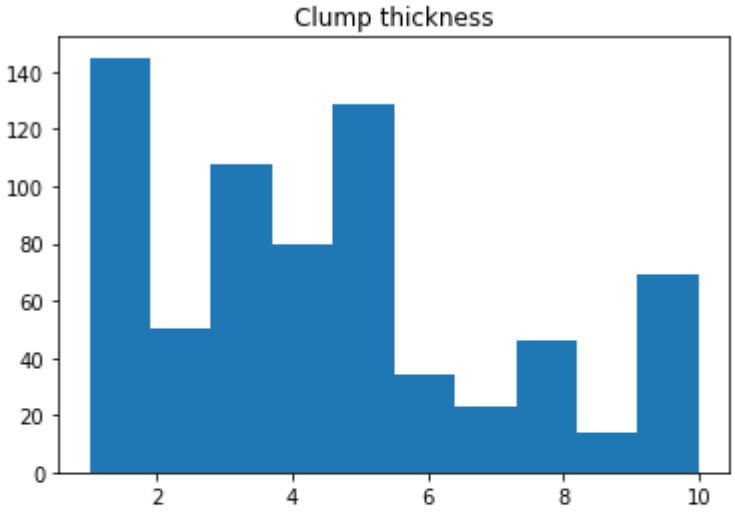
Dane we wszystkich zbiorach rozpinają się na przedziale od 1 do 10, wypełniając całą dostępną skalę. **To oznacza, że w badanej próbce występują różne poziomy tych parametrów, a skalę wartości w pełni wykorzystano.**

Dla (1) średnia i mediana są dość blisko siebie (0.4 różnicy) w porównaniu do (2) i (3), które różnią się aż o ~2 wartości. Dla wszystkich przypadków kwartyle od 0.25 do 0.75 mają rozpiętość 5 wartości, ale dodatkowo (1) jest przesunięty o jedną wartość wyżej (tzn. od 2 do 6). **To sugeruje, że większość wartości dla tego parametru jest skupiona w wyższych wartościach, co może wskazywać na większą obecność grudek o większej grubości w badanych próbkach.**

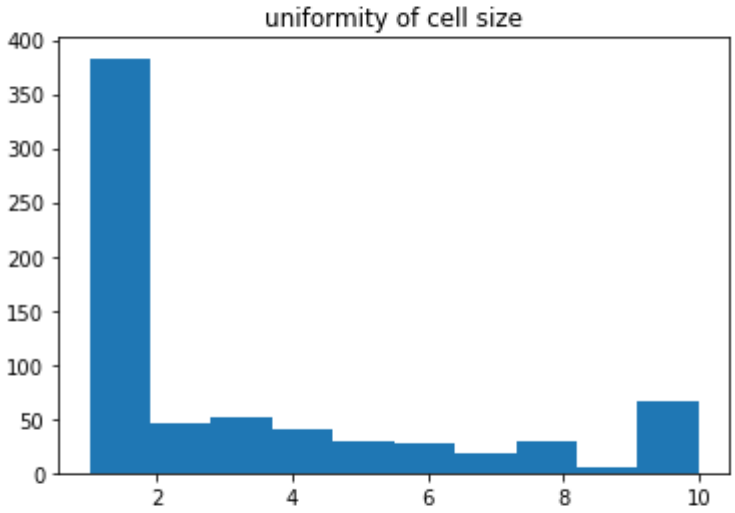
Można zauważyć również, że dla (2) i (3) mediana jest taka sama i jednocześnie kwartyle również są takie same, w przeciwieństwie do (1). **Może to oznaczać, że te dwa parametry są silnie skorelowane i mogą dostarczać podobnych informacji na temat badanych komórek.**

Jeśli dwie grupy mają podobną średnią wartość, ale jedna z grup ma większe standardowe odchylenie, oznacza to, że dane w tej grupie są bardziej zmiennościowe, co może oznaczać, że dane w zbiorze mają większe rozproszenie wokół średniej wartości. Jeśli zestaw danych jest bardziej zmiennościowy, oznacza to, że wartości w nim są bardziej zróżnicowane i rozproszone, a odległość między poszczególnymi wartościami może być większa. **Może to wskazywać na większą zróżnicowanie wielkości i kształtu komórek w próbce danych.**

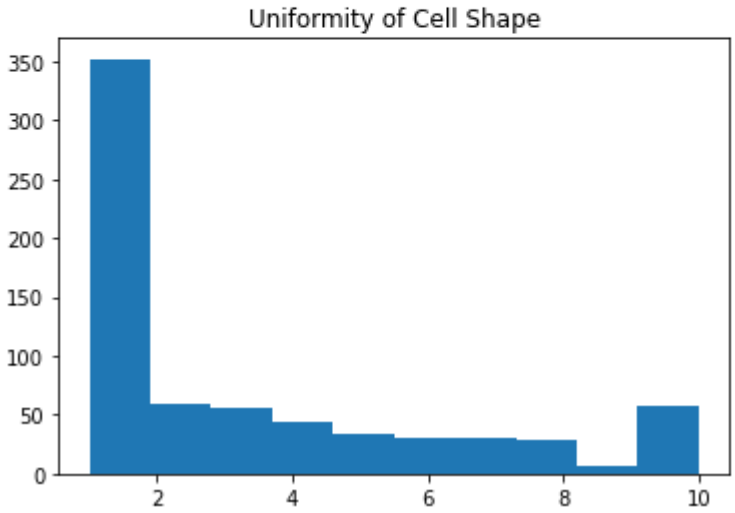
4 Zrobić wykresy dla trzech wybranych zmiennych (wykres liniowy, histogram, pudełkowy). Przedstawić ich analizę



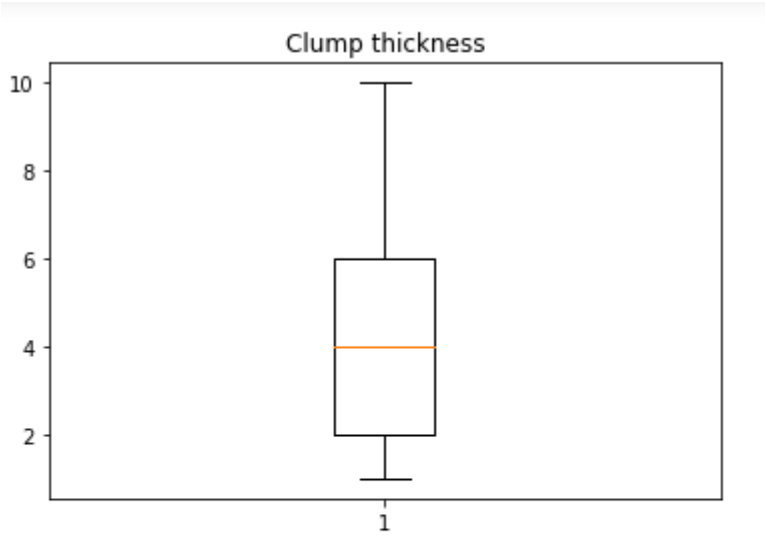
W Clump thickness można zauważyć, że większość danych znajduje się w dolnej części wartości. Najniższa zauważalna wartość to 9.

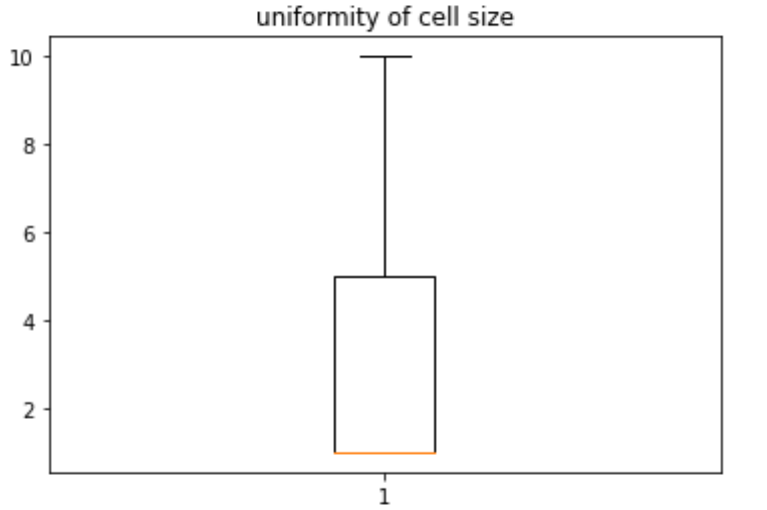


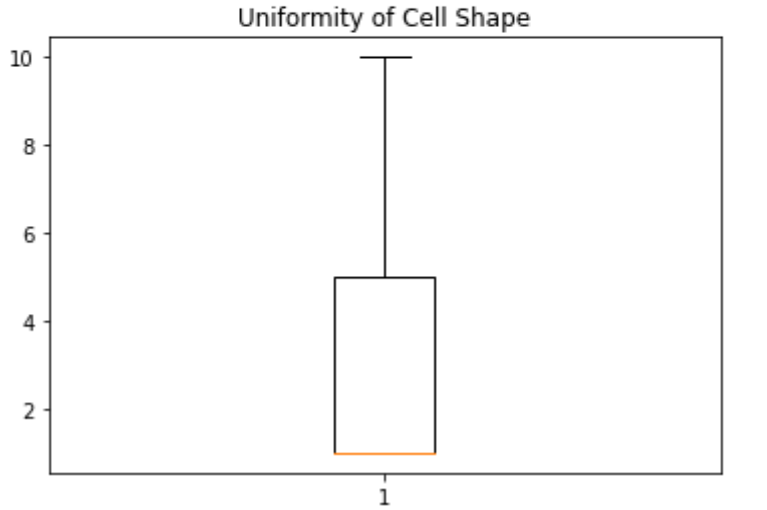
W Uniformity of cell size widać, że większość danych (ponad 350) ma wartość jeden. Dostrzec można również prawie całkowity brak wartości 9 na histogramie.



W Uniformity of cell shape widać, że większość danych (ponad 300) ma wartość jeden. Dostrzec można również prawie całkowity brak wartości 9 na histogramie.

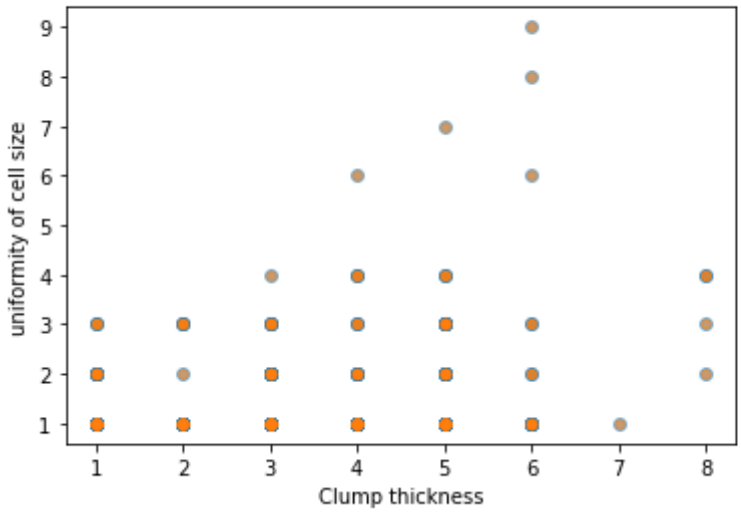






Na wykresie pudełkowym dla Uniformity of cell size i Uniformity of cell shape widać, że mediana przybiera wartość jeden. W konsekwencji co najmniej połowa danych ma wartość jeden.

Korelacja Clump thickness i uniformity of cell size dla Niegorźnych

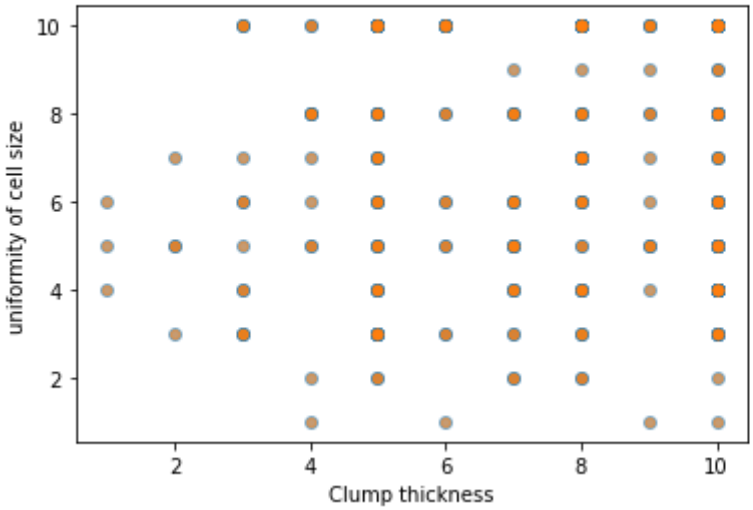


Widać, że dla klasy Niegroźne, Uniformity of cell size o wartości jeden występuje silna korelacja pomiędzy clump thickness w przedziale (1,6). Więc w przyszłości dostrzegając taką zależność można sugerować, że badana próbka ma przynależność do klasy Niegroźne.

Na Histogramach można zauważyć jednak, że większość danych ma wartość jeden, przez co korelacja ta nie jest miarodajna.

W Niegroźnej klasie można zaobserwować dla Clump thickness całkowity brak korelacji dla wartości 9 i 10.

Korelacja Clump thickness i uniformity of cell size dla Złośliwych



W tym przypadku dla klasy Złośliwej, można zauważyć silną korelację w punktach 5, 7, 8 i 10 dla Clump thickness pomiędzy uniformity of cell size kolejno w przedziałach (10, 8-2), (8, 6-2), (10,8-2), (10-3). O ile pojedyncze korelacje na próbkach mogą budzić w przyszłości obawy, tak zbiory powyżej przedstawionych korelacji w przyszłych badaniach mogą stanowić silną sugestię dla diagnozy próbki jako Złośliwej.

WNIOSKI

Wyniki analizy można podsumować w następujących punktach:

- Badany zbiór danych dotyczył analizy parametrów związanych z rakiem.
- Zbiór zawiera dwie klasy: Niegroźną (457 próbek) i Złośliwą (241 próbek).
- Różnica w liczbie próbek między klasami może wpływać na wyniki uczenia maszynowego. -W przypadku nierównomiernego rozkładu danych, model może wykazywać skrzywienie i być bardziej skłonny do przewidywania dominującej klasy.
- Wykryto brakujące dane tylko w jednej kolumnie ("Bare Nuclei").
- Nierównomierne rozłożenie danych może prowadzić do niskiej dokładności modelu oraz trudności w wykrywaniu rzadkich zdarzeń.
- Analiza trzech wybranych zmiennych wykazała różnice w średnich, medianach i rozproszeniu wartości.
- Wykresy liniowy, histogram i pudełkowy dla tych zmiennych mogą dostarczyć dodatkowych informacji o rozkładzie danych.
- Podobne kwartyle pozwoliły zauważyć, że między dwoma podzbiorami może występować korelacja co warto sprawdzać w przyszłych przypadkach.

Podsumowując, ważne jest uwzględnienie nierównomiernego rozkładu danych oraz związanych z tym potencjalnych skutków podczas analizy i modelowania danych.

Należy również skorygować brakujące wartości i zastosować odpowiednie techniki równoważenia danych, aby uniknąć skrzywienia modelu i zachować pełność informacji w zbiorze danych.