1. **Zadanie**

W agencji nieruchomości „Tani dom” wyróżniono trzy cechy charakteryzujące nieruchomości przeznaczone do wynajmu są to: Duże, Tanie, Uzbrojone.

W zależności od wielkości, ceny i uzbrojenia mieszkania w media do cech Duże (D), Tanie (T), Uzbrojone (U) przypisywane są wartości wag  od 1 do 5. Agencja wykonała badania ankietowe na wybranych klientach którzy zaliczali mieszkania na podstawie wag do dwóch grup preferencji. Gdzie grupa pierwsza oznacza mieszkania najbardziej poszukiwane.

Część danych z badań ankietowych pokazuje poniższa tabela:

Obraz zawierający biały, meble, szafka

Opis wygenerowany automatycznie

Rys. 1. Fragment danych ankietowych.

Zadania do wykonania:

A.

1. Zbudować drzewo klasyfikacyjne przystosowane do klasyfikacji mieszkań
2. Sformułować wnioski.

B.

1. Wygenerować dane losowych preferencji dla 100 przypadków mieszkań dla cech: Duże (D), Tanie (T), Uzbrojone (U),  Blisko Centrum (BC). Każda z tych cech może przybrać wartości wag od 1 do 8. Należy wykorzystać funkcję LOS().
2. **Zbudować funkcję klasyfikującą** każdą ze 100 nieruchomości do grupy pierwszej lub drugiej.
3. Zbudować drzewo klasyfikacyjne dostosowane do klasyfikacji mieszkań do grupy pierwszej lub drugiej.
4. Sformułować wnioski.

**C.**

1. **Wykonać zadanie z podpunktu B. wykorzystując inne metody (bagging random forest, svm).**
2. **Sformułować wnioski.**
3. **Sposób rozwiązania**

Do rozwiązania zadania wykorzystano biblioteki R *tree, ipred, randomForest, e1071* oraz *rattle*, które pozwalają na budowę modeli klasyfikacyjnych.

1. **Implementacja w środowisku R**

Zgodnie z założeniami implementację podzielono na cztery części. Każda z nich obejmowała budowę modeli wykorzystując różne metody klasyfikacji metody.

1. ***Podpunkt A***

Pierwsza część zadania (podpunkt A) został rozwiązany podczas zajęć, dlatego ograniczono się jedynie do umieszczenia delikatnie skorygowanego kodu źródłowego.

Obraz zawierający tekst

Opis wygenerowany automatycznie

1. ***Podpunkt B***

W celu rozwiązania zadania B należało najpierw wygenerować losowe dane   
w programie Excel. Do generowania danych wykorzystano formułę LOS(), a do klasyfikacji mieszkań zbudowano funkcję o zależnościach przedstawionych poniżej.

Obraz zawierający stół

Opis wygenerowany automatycznie

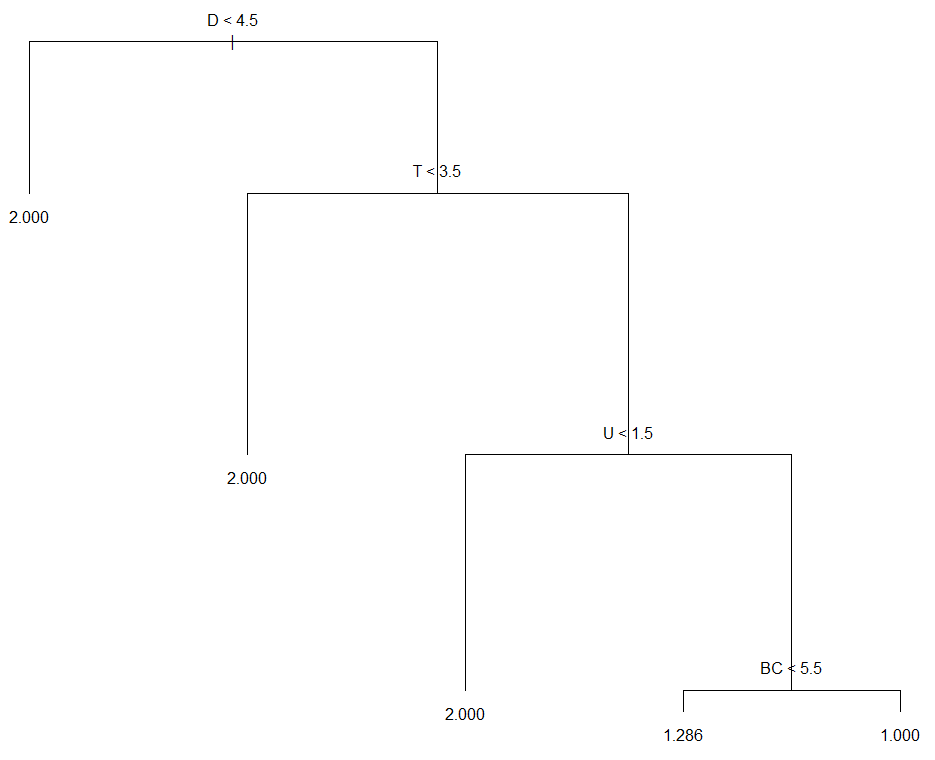


Wygenerowany w ten sposób zbiór danych zapisano do osobnego pliku .csv,   
a następnie zbudowano na jego podstawie dwa drzewa klasyfikacyjne – po jednym   
z wykorzystaniem jednej z dwu bibliotek do tworzenia drzew klasyfikacyjnych.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Wykonanie powyższego kodu tworzy dwa drzewa klasyfikacyjne dla wygenerowanego zbioru danych. Pierwsze z nich, wykorzystujące bibliotekę *tree* przedstawiono poniżej.



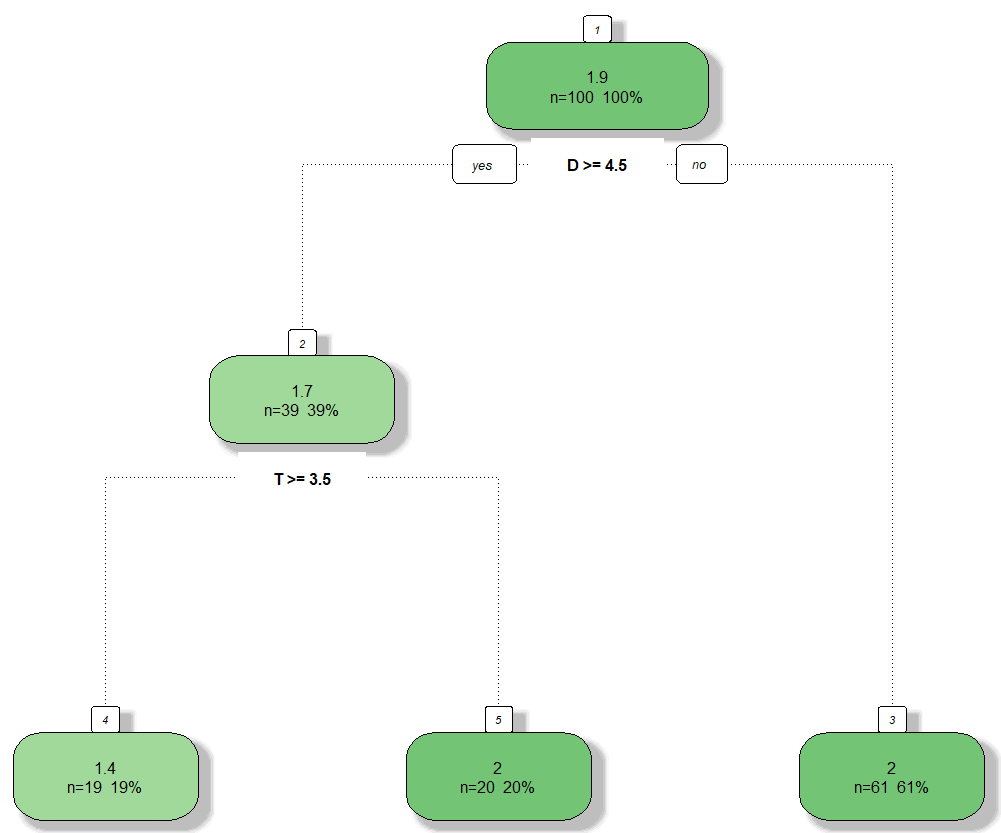
Model uzyskany dla tego drzewa charakteryzuje się następującymi parametrami:

Obraz zawierający tekst

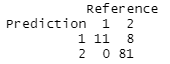
Opis wygenerowany automatycznie

Jak widać dla wygenerowanych danych, przy wykorzystaniu biblioteki *tree* udało się osiągnąć bardzo dobry model. Tylko 2 mieszkania zostały źle przypisane, co równoznaczne jest z osiągnięciem dokładności na poziomie 98%.

Druga z wykorzystanych bibliotek pozwoliła dla tego samego zbioru danych wygenerować następujące drzewo klasyfikacyjne.



Model dla tego drzewa charakteryzuje się następującymi parametrami:



Nietrudno zauważyć, że model ten jest nieco gorszy od poprzedniego. Pozwolił na osiągnięcie dokładności 92%.

1. **Inne metody (Podpunkt C)**

Przed przystąpieniem do implementacji wysnuto hipotezę, że wykorzystanie jednej z przedstawionych w tym podpunkcie metod pozwoli na osiągnięcie lepszego modelu niż w przypadku drzew klasyfikacyjnych.

Inne metody sprawdzano na dodatkowym zestawie danych stanowiących swego rodzaju walidację.

* 1. Bagging

Wykorzystanie metody pozwoliło na osiągnięcie następujących wyników nauki:

Obraz zawierający tekst, paragon

Opis wygenerowany automatycznie

Dla zbioru walidacyjnego osiągnięto następujące wyniki:

Obraz zawierający tekst, paragon

Opis wygenerowany automatycznie

Jak nietrudno zauważyć model jest porównywalny z drzewami klasyfikacyjnymi zbudowanymi w pierwszej części zadania. Dla zbioru walidacyjnego osiągnięto dokładność na poziomie 95%.

* 1. Random Forest

Wyniki uczenia:

Obraz zawierający tekst, paragon

Opis wygenerowany automatycznie

Dla zbioru walidacyjnego osiągnięto następujące wyniki:

Obraz zawierający tekst, paragon

Opis wygenerowany automatycznie

Metoda Random Forest uzyskała 100% dokładność dla zbioru walidacyjnego!

* 1. SVM

Wyniki uczenia:

**Obraz zawierający tekst, paragon

Opis wygenerowany automatycznie**

Dla zbioru walidacyjnego osiągnięto następujące wyniki:

**Obraz zawierający tekst, paragon

Opis wygenerowany automatycznie**

Wyniki dla SVM są porównywalne z drzewami klasyfikacyjnymi zbudowanymi   
w pierwszej części zadania. Dla zbioru walidacyjnego osiągnięto dokładność na poziomie 95%.

* 1. **Wnioski**

Budowanie modeli klasyfikacyjnych w środowisku R jest bardzo przejrzyste   
i proste w obsłudze. W trakcie prac najdokładniejsze modele udawało się uzyskać dla modeli **Random Forest**, które niezależnie od asymetrii danych osiągały dokładność na poziomie 100%.

Warto nadmienić, że w zależności od wygenerowanych danych, mięliśmy do czynienia z mniejszą, lub większą asymetrią danych, co wpłynęło na jakość uzyskanych modeli. Drzewa klasyfikacyjne, bagging oraz SVM wypadają podobnie, chociaż gorzej od **Random Forest**!