Dokumentacja

Wnioski z raportu automatycznego

Ogólne statystyki zbioru danych:

Zbiór składa się z: 16 zmiennych (5 liczbowych i 11 kategorycznych) i 4024 obserwacji.

Rozkłady zmiennych:

- W kolumnie „Age” mamy duży rozrzut danych co pokazuje że mamy informację o ludziach z różnych grup wiekowych. Na histogramie widać asymetrię w kierunku wyższych wartości, a średnia wieku wynosi 53,97 co może wskazywać na fakt że rak piersi dosięga częściej osoby starsze.
- W kolumnie „Race” Większość danych pochodzi od osób z kategorii "White", co może mieć wpływ na reprezentatywność wyników analizy.
- Patrząc na zmienną Status można zobaczyć że większość pacjentów jest wciąż żywa (3408).

Brakujące wartości i duplikaty:

- Brak brakujących wartości w danych.
- Jeden zduplikowany wiersz.

Korelacje: Wykryto znaczące korelacje między zmiennymi:

- N Stage i 6th Stage: 0.88
- N Stage i Regional Node Positive: 0.84
- T Stage i Tumor Size: 0.81
- 6th Stage i Regional Node Positive: 0.77
- T Stage i 6th Stage: 0.61
- Survival Months i Status: -0.48

Wartości odstające występują w kolumnach:

- Tumor Size – większość wartości poniżej 50, mimo to istnieją wartości maksymalne równe 140.
- Regional Node Positive – dane skupione poniżej 20, a wartości odstające wynoszą nawet 46.

- Regional Node Examined – również skumulowane w niższych wartościach, średnia to 14,36, a wartości odstające sięgają nawet do 61.

Transformacje kategorii:

- W kolumnie „Grade” występują wartości: 1, 2, 3 i "Anaplastic; Grade IV" więc pojawia się tu pewna niespójność której można się pozbyć.

Niezbalansowane zmienne:

- Kolumny A Stage i Estrogen Status są niezbalansowane, a po sprawdzeniu ważności cech okazało się że mają kolejno ważność 0.0040 i 0.0161 więc mają bardzo niski wpływ na model.

Po analizie raportu stworzono skrypt clean_data.ipynb, który:

- Usuwa 1 duplikat.
- Usuwa wartości odstające w kolumnach:

  Tumor Size – Usunięto 222 wierszy.

  Regional Node Positive - Usunięto 276 wierszy.

  Regional Node Examined - Usunięto 65 wierszy.
- Zmienia wartość "Anaplastic; Grade IV" na liczbę "4" i konwertuje całą kolumnę na numeryczną.
- Usuwa kolumny A Stage i Estrogen Status.
- Przygotowuje dane do modelowania.

Wybrane Narzędzie AutoML:

TPOT (Tree-based Pipeline Optimization Tool) - narzędzie AutoML, które automatyzuje proces projektowania, optymalizacji i wybiera najlepsze modele, co pozwala na znaczną oszczędność czasu i zasobów podczas eksploracji różnych konfiguracji modeli. Dzięki TPOT użytkownicy mogą skupić się na analizie wyników, zamiast ręcznie dostrajać hiperparametry czy budować złożone potoki przetwarzania danych.

Wady:

- Brak wsparcia dla głębokiego uczenia.
- Konieczność ręcznego przetwarzania braków danych.

Wyniki TPOT

Generation 1 - Current best internal CV score: 0.9124691147652723

Generation 2 - Current best internal CV score: 0.9124691147652723

Generation 3 - Current best internal CV score: 0.9124691147652723

Generation 4 - Current best internal CV score: 0.9124708187782227

Generation 5 - Current best internal CV score: 0.9141220073272557

Pipeline 1:

Name: RandomForestClassifier(bootstrap=False, criterion=entropy, max_features=0.9, min_samples_leaf=16, min_samples_split=3, n_estimators=100)

Pipeline 2:

Name: DecisionTreeClassifier(criterion=entropy, DecisionTreeClassifier__max_depth=4, min_samples_leaf=6, min_samples_split=2)

Pipeline 3:

Name: ExtraTreesClassifier(bootstrap=True, criterion=entropy, max_features=0.8, min_samples_leaf=3, min_samples_split=12, n_estimators=100)

Opis modeli wybranych przez TPOT

1. **RandomForestClassifier –** Model oparty na wielu niezależnych drzewach decyzyjnych, które uczą się na losowych próbkach danych. Dzięki tej strategii jest odporny na overfitting, co czyni go wysoce efektywnym w różnych zastosowaniach.
    - Wynik: Internal CV Score = 0.9141.
2. **DecisionTreeClassifier -** Model działa poprzez iteracyjne dzielenie danych na podzbiory za pomocą warunków logicznych. Jego kluczową zaletą jest prostota oraz łatwość interpretacji wyników, co czyni go idealnym wyborem w zastosowaniach wymagających transparentności i zrozumiałości działania modelu.
    - Wynik: Internal CV Score = 0.9129.
3. **ExtraTreesClassifier -** Model zbliżony do RandomForestClassifier, w którym wprowadzono dodatkową losowość podczas wyboru granic podziału w węzłach drzew. Dzięki temu proces uczenia jest szybszy, a model staje się bardziej wydajny obliczeniowo i mniej podatny na przeuczenie.
    - Wynik: Internal CV Score = 0.9125.

Wybrany model

Postanowiłem wybrać **RandomForestClassifier** ponieważ osiągnął najwyższą skuteczność wynoszącą 0.9141. Model ten jest odporny na przeuczenie dzięki losowemu wyborowi próbek i cech, co pozwala na lepszą generalizację wyników. Dodatkowo, mechanizm łączenia wielu drzew decyzyjnych sprawia, że model jest elastyczny i skuteczny w analizie nieliniowych zależności w danych.

Najlepsze parametry jakie wybrał TPOT to:

- bootstrap=False
- criterion=entropy
- max_features=0.9
- min_samples_leaf=16
- min_samples_split=3
- n_estimators=100