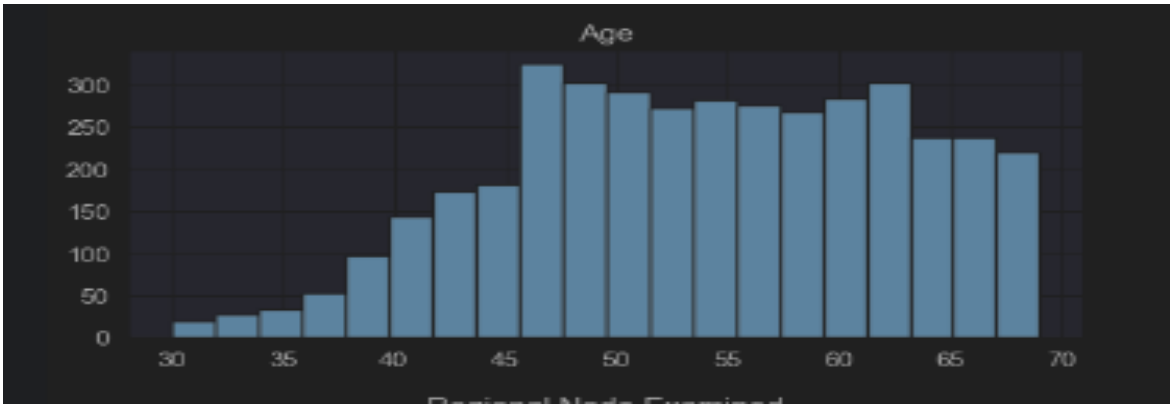
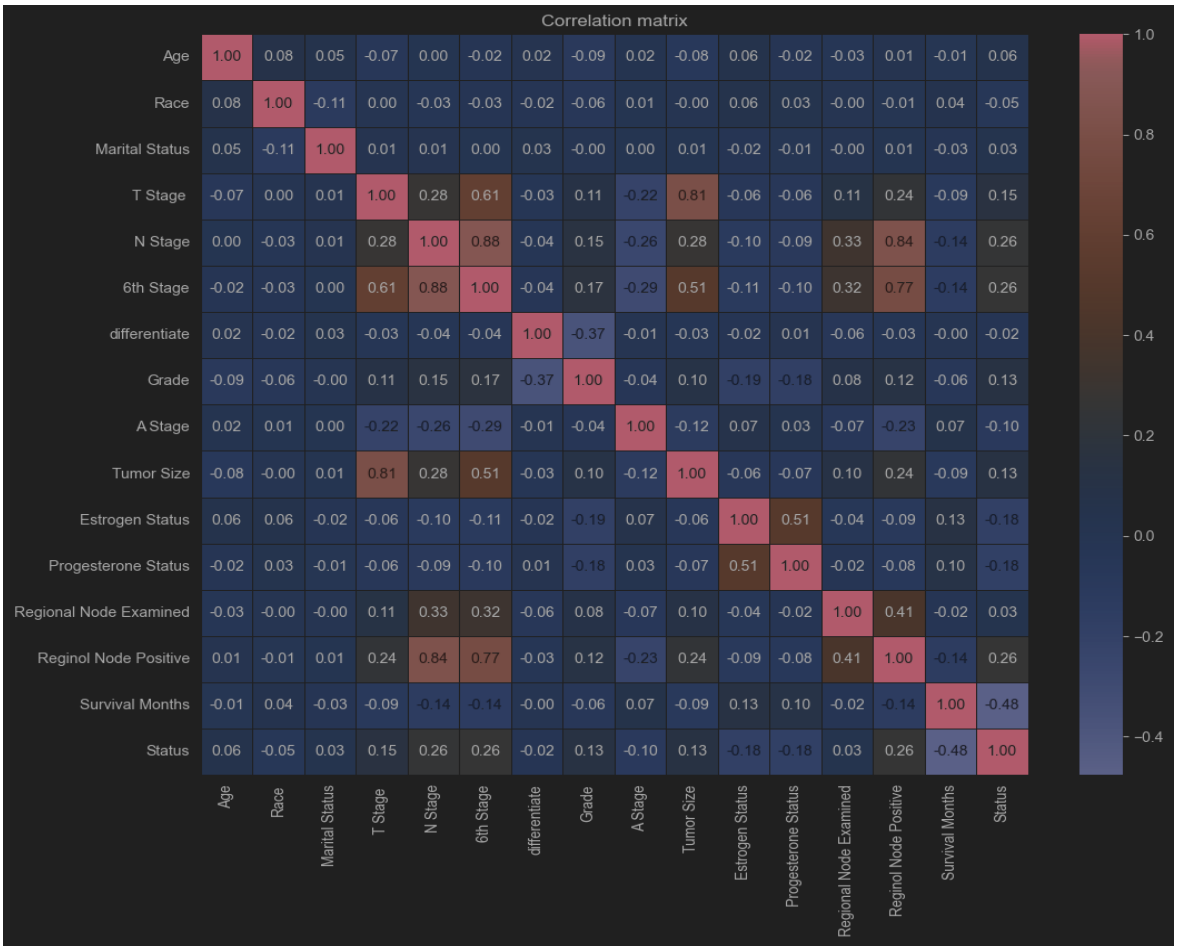
Raport

Podsumowanie wyników analizy danych

W wyniku analizy danych zauważono, że większość danych pochodzi od osób w starszym wieku, co może wskazywać na wyższe ryzyko zachorowania na raka piersi wśród starszych osób. Wystąpiły również korelacje między różnymi zmiennymi, szczególnie związane z etapami choroby i rozmiarem guza.





Przetwarzanie danych

W procesie przetwarzania danych dla cech liczbowych zastosowano metodę skalowania przy użyciu **StandardScaler**, która standaryzuje dane, aby miały średnią 0 i odchylenie standardowe 1. Cechy kategoryczne zostały natomiast zakodowane za pomocą metody **One-Hot Encoding**. Procesy te zostały zintegrowane w jeden etap przy pomocy **ColumnTransformer**, który umożliwia równoczesne przetwarzanie różnych typów danych.

Wyniki Modelu

Do budowy modelu klasyfikacyjnego wykorzystano algorytm **RandomForestClassifier** z następującymi, zoptymalizowanymi hiperparametrami:

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

Po przeprowadzeniu tego etapu dane zostały podzielone na zbiór treningowy (70%) oraz testowy (30%). Model oceniano przy pomocy standardowych miar klasyfikacyjnych, uzyskując następujące wyniki:

- **Dokładność**: 0.9171
- **Precyzja**: 0.9295
- **Czułość**: 0.9790
- **Miara F1**: 0.9536

Stworzono również Raport klasyfikacyjny:

	precision	recall	f1-score	support
Alive	0.93	0.98	0.95	903
Dead	0.78	0.50	0.61	135
accuracy			0.92	1038
macro avg	0.86	0.74	0.78	1038
weighted avg	0.91	0.92	0.91	1038

Wnioski

Model wykazuje bardzo dobre wyniki, szczególnie w zakresie **czułości**, co wskazuje na jego zdolność do wykrywania pozytywnych przypadków (np. klas pozytywnych w problemie klasyfikacji). Wysoka **precyzja** (0.9295) sugeruje, że model rzadko myli klasy pozytywne z negatywnymi, co jest szczególnie ważne w przypadku, gdy koszt błędów fałszywych alarmów jest wysoki. **Dokładność** na poziomie 91.71% również potwierdza skuteczność modelu. Dodatkowo, wysoka wartość miary **F1** (0.9536) wskazuje na dobry kompromis między precyzją a czułością, co czyni model dobrze zbalansowanym.

Dalsze kierunki rozwoju modelu

Eksperymenty z hiperparametrami

Aby poprawić wydajność modelu, warto przeprowadzić bardziej zaawansowaną optymalizację hiperparametrów. Techniki takie jak Grid Search lub Random Search mogą pomóc w znalezieniu optymalnych wartości parametrów, takich jak liczba drzew, maksymalna liczba cech czy minimalna liczba próbek w liściu.

Inżynieria cech (Feature Engineering)

Ze względu na wysoką korelacją między dużą ilością zmiennych w danych stworzenie nowych cech pochodnych może zwiększyć moc predykcyjną modelu.

Walidacja krzyżowa

W celu dokładniejszej oceny zdolności modelu do generalizacji warto zastosować k-krotną walidację krzyżową. Taka metoda pozwala lepiej zrozumieć, jak model radzi sobie na różnych podzbiorach danych, co zwiększa pewność, że uzyskane wyniki są stabilne i reprezentatywne dla całego zbioru danych.