Klasyfikacja podgrupy raka na podstawie proteomów

Maja Andrzejczuk
Piotr Bielecki
Michał Iwicki
Julia Przybytniowska

June 2023

1 Wstęp

2 Wykorzystane dane

Wykorzystane zostały następujące zestawy danych:

- 77_cancer_proteomes_CPTAC_itraq - główny zestaw danych, który zawiera informacje dotyczące 77 próbek raka piersi. Próbki te zostały wygenerowane przez Clinical Proteomic Tumor Analysis Consortium (NCI/NIH), dla każdej z nich zarejestrowano wartości ekspresji dla około 12 000 białek. Te dane proteomowe stanowią podstawę analizy i klasyfikacji próbek.
- clinical_data_breast_cancer - służy do dopasowania identyfikatorów próbek w głównym pliku proteomów. Zawiera również dodatkowe informacje dotyczące klasyfikacji nowotworów dla poszczególnych próbek, które zostały określone za pomocą różnych metod. Może zawierać informacje takie jak stadium nowotworu, stopień złośliwości, reakcja na terapię, dane kliniczne pacjentów itp.
- PAM50_proteins - zawiera listę genów i białek używanych przez system klasyfikacji PAM50. PAM50 to metoda klasyfikacji molekularnej, która pomaga w identyfikacji podtypów raka piersi na podstawie profilu ekspresji genów. Zawiera identyfikatory białek, które można dopasować do identyfikatorów znajdujących się w głównym zestawie danych ekspresji białek.

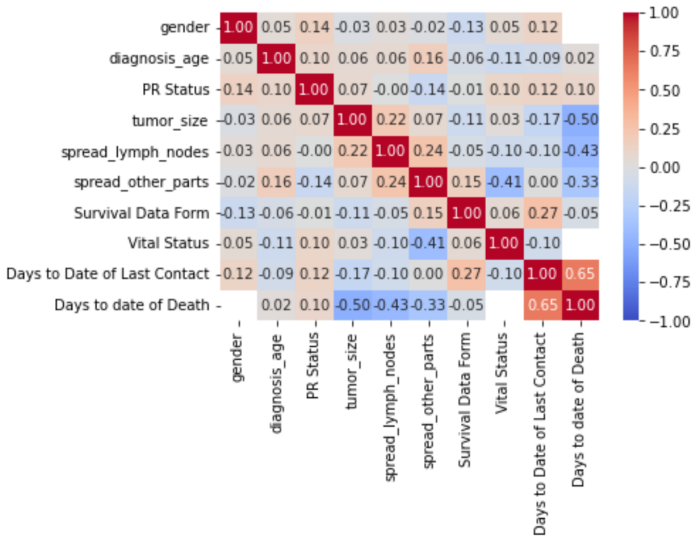
3 Inżynieria danych

W celu przygotowania danych do modelu, zastosowano kilka kroków preprocessingu. Na początku, z ramki danych o nazwie "clinical_data_breast_cancer" usunięto kolumny, które zawierały powtarzające się informacje w innych kolumnach, czyli wysoko skorelowane zmienne. Takie redundantne informacje nie wnosiły dodatkowej wartości do analizy. Usunięto również kolumny, które miały tylko jedną unikatową wartość, ponieważ nie przyczyniały się do zrozumienia różnic między próbkami.

Następnie przeprowadzono wstępne zakodowanie niektórych zmiennych. Zmienna płci oraz ta opisująca przerzuty do innych organów zostały zakodowane, aby reprezentowały wartości binarne, natomiast stopień raka oraz ten opisujący zakażenia układów linfatycznych zostały zakodowane jako zmienna liczbowe, kategoryczne, aby uwzględnić różne poziomy. Zmienna opisująca wiek zdiagnozowanego pacjenta została zakodowana na 3 kategorie ('young', 'middle', 'elderly').

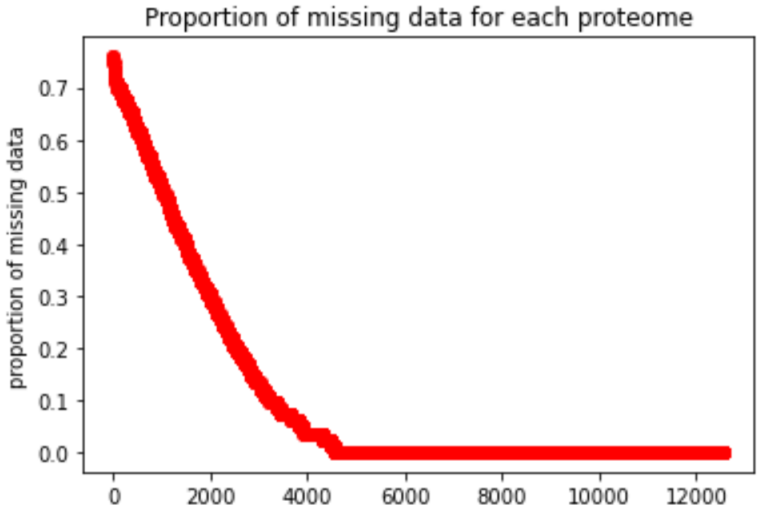
Kolejnym podjętym krokiem było sprawdzenie korelacji między zmiennymi w celu ustalenia, czy istnieją dodatkowe kolumny, które powinny zostać usunięte. Prezentowały się one następująco:

Następnie skoncentrowaliśmy się na analizie 77_cancer_proteomes_CPTAC_itraq. Okazało się, że duża część tego zestawu danych była brakująca, co mogłoby mieć negatywny wpływ na dokładność



Rysunek 1: Mapa korelacji wybranych zmiennych.

analizy i modelowania. Zatem aby lepiej zrozumieć rozmiar problemu, stworzyliśmy wykres, prezentujący jaką część kolumn stanowią wartości NaN.



Rysunek 2: Caption

Widoczne jest że ponad 2,5 tysięcy protomów posiada wiecej niż 20% brakujących danych w swojej kolumnie. Uznaliśmy, że usunięcie 2.5 tysiąca kolumn nie jest zbyt duża stratą, a braki w reszcie kolumn wypełniliśmy wartością średnią.

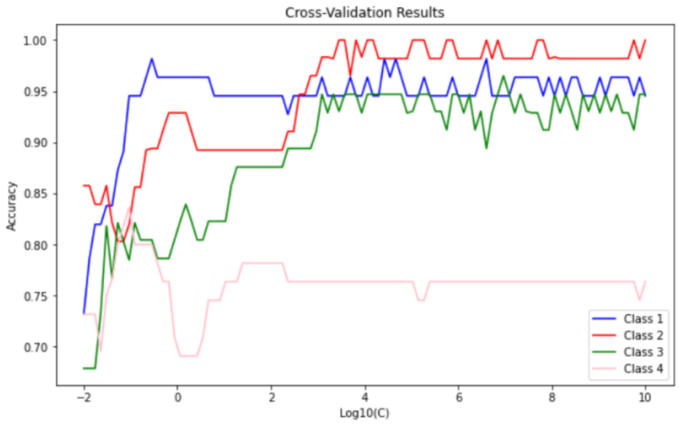
Ostatnim podjętym krokiem w celu przygotowania danych pod klasyfikację było złączenie wszystkich plików formatu 'csv' oraz zastosowanie na niektórych kolumnach One Hot Encoding.

4 Wykorzytane modele

W naszym rozwiązaniu wykorzystaliśmy trzy różne modele do klasyfikacji typów nowotworów na podstawie danych wejściowych.

4.1 Logistic Regression (Regresja Logistyczna)

Wykorzystano model *LogisticRegressionCV* z pakietu *scikit-learn*. Zastosowano walidację krzyżową (cross-validation) z parametrem *cv=5*, co oznaczało podział danych na 5 podzbiorów i iteracyjne trenowanie modelu na 4 podzbiorach, a następnie testowanie na pozostałym podzbiorze. Wykorzystano regularyzację L1 (*penalty='l1'*) i solver *'liblinear'*. Przeprowadzono przeszukiwanie siatki (grid search) po wartościach parametru C, który reprezentuje odwrotność siły regularyzacji. Wybrano najlepszą wartość parametru C na podstawie wyników walidacji krzyżowej. Obliczono dokładność (*accuracy*) modelu na podstawie wyników walidacji krzyżowej dla różnych wartości parametru C. Wykres pokazuje zależność między logarytmem wartości C a dokładnością dla różnych klas nowotworów.



Rysunek 3: Wykres modelu Regresji Logistycznej

4.2 K-Nearest Neighbors (K-najbliższych sąsiadów)

Wykorzystano model *KNeighborsClassifier* z pakietu *scikit-learn*. Zastosowano walidację krzyżową, dzieląc dane na zbiory treningowe i testowe w różnych iteracjach. Przetestowano model dla różnych wartości parametru k (liczba sąsiadów). Obliczono średnią dokładność (*accuracy*) oraz odchylenie standardowe dla każdego testowanego k. Wyświetlono wyniki dla każdej wartości k.

4.3 Support Vector Machine

Wykorzystano model *SVC* z pakietu *scikit-learn*. Zastosowano walidację krzyżową, dzieląc dane na zbiory treningowe i testowe w różnych iteracjach. Wykorzystano kernel radialnej funkcji bazowej (RBF) poprzez ustawienie parametru *kernel="rbf"*. Obliczono średnią dokładność (*accuracy*) oraz odchylenie standardowe na podstawie wyników walidacji krzyżowej.

5 Wybór najistotniejszych białek

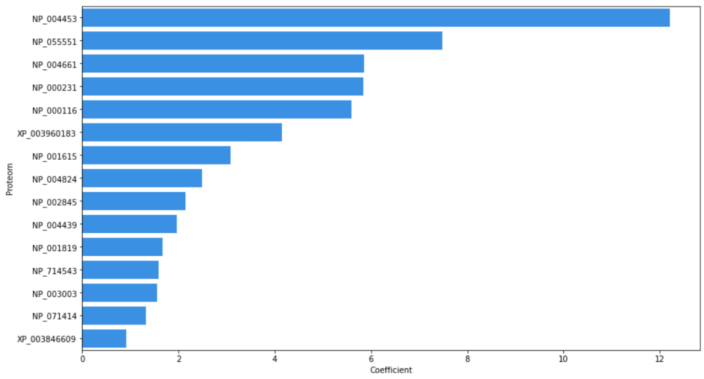
Najlepsze wyniki dla zbiorów testowych osiągnęliśmy przy użyciu modelu Regresji Logistycznej, dlatego też dalsze działania kontynuowaliśmy opierając się na tym modelu.

5.1 Selekcja białek

W celu dokonania selekcji białek zastosowaliśmy metodę RFE (Recursive Feature Elimination). Metoda ta umożliwia wybór zmiennych mających największy wpływ na dany model. Jej działanie polega na rekurencyjnym redukowaniu liczby zmiennych poprzez obliczanie ważności cech (feature importance) lub współczynnika (coefficient) za pomocą kroswalidacji, a następnie wybieraniu najistotniejszych kolumn.

Warto zaznaczyć, że metoda RFE jest niedeterministyczna, co oznacza, że zmienne o niewielkim wpływie mogą różnić się w zależności od konkretnej iteracji. W naszym przypadku, dokonując selekcji białek. Choć metoda RFE jest teoretycznie niedeterministyczna i może prowadzić do różnych wyników w różnych iteracjach, w naszym przypadku zauważyliśmy, że podczas wielokrotnych wywołań często uzyskiwaliśmy te same białka jako najważniejsze. Te powtarzające się wyniki sugerują, że istnieją białka o znaczącym wpływie na nasz model i które są istotne dla badania.

Celem naszego badania było wyselekcjonowanie jak najmniejszej liczby białek, które jednocześnie zapewniałyby dobre wyniki. Dzięki temu moglibyśmy opracować badanie, które byłoby efektywne i przydatne dla pacjentów. Przeprowadziliśmy piętnastokrotne generowanie białek, a następnie na tej podstawie wybraliśmy końcowy zestaw białek.



Rysunek 4: Najważniejsze białka

Badając uzyskane białka, przeprowadziliśmy analizę genów, które są za ich powstawanie odpowiedzialne. Dzięki temu gromadzimy wartościowe informacje dotyczące mechanizmów regulacyjnych, interakcji molekularnych oraz potencjalnych funkcji tych białek w organizmie. Wspomniane nazwy możemy odczytac za pomogą poniższej tabeli:

	Protein	Gene Symbol	Coefficient
0	NP_004453	FDFT1	11.662615
1	NP_055551	KIAA0101	7.121585
2	NP_000231	MAOA	5.580493
3	NP_004661	PAPSS2	5.530496
4	NP_000116	ESR1	5.267939
5	XP_003960183	NaN	3.891289
6	NP_001615	AIM1	2.917412
7	NP_004824	AIM2	2.373186
8	NP_002845	PVALB	2.032070
9	NP_004439	ERBB2	1.909760
10	NP_001819	CLC	1.603851
11	NP_003003	SFRP1	1.545582
12	NP_714543	GSTA5	1.545398
13	NP_071414	CLSTN2	1.206695
14	XP_003846609	NaN	0.912354

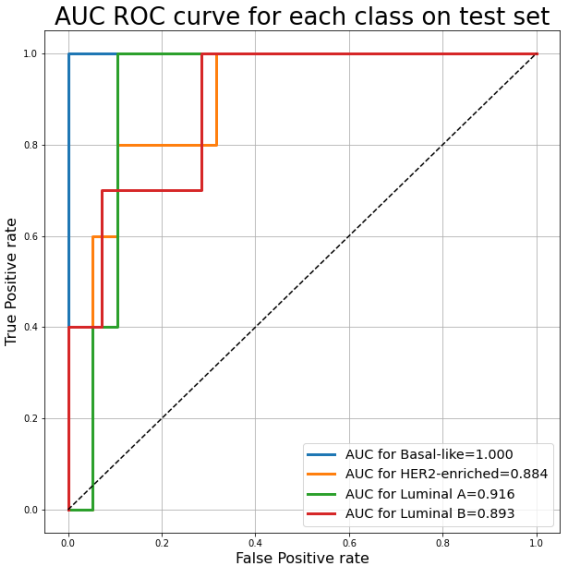
Rysunek 5: Najważniejsze białka oraz odpowiadające im geny

6 Analiza wyników

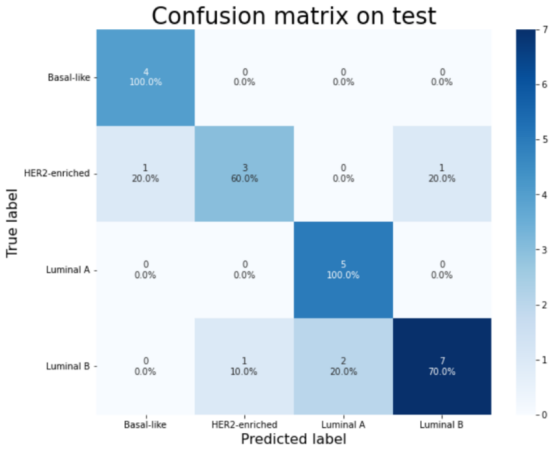
Aby ocenić skuteczność naszego modelu, przeanalizowaliśmy różne metryki, które dostarczają informacji na temat jego wydajności. Poniżej przedstawiamy wyniki naszych modeli wraz z odpowiednimi metrykami:

Metryka	Wynik
F1 score	0.785
Roc_auc_score	0.922
Test accuracy	0.792
Crossval score	0.818
Crossval std	0.9

F1 score mierzy zharmonizowane średnie między precyzją i czułością, gdzie wartość 1 oznacza doskonałą klasyfikację. Roc_auc_score odzwierciedla zdolność modelu do rozróżnienia klas, gdzie wartość 1 oznacza doskonałą zdolność do klasyfikacji. Test accuracy to dokładność modelu w przewidywaniu klas na zbiorze testowym. Crossval score oznacza wynik walidacji krzyżowej, która ocenia skuteczność modelu na różnych podzbiorach danych. Wreszcie, Crossval std to odchylenie standardowe wyników walidacji krzyżowej, które mierzy stabilność modelu.



Rysunek 6: Krzywa ROC



Rysunek 7: Macierz pomyłek

Wartość AUC to miara, która ocenia zdolność modelu do rozróżnienia między klasami. Im bliżej wartości 1, tym lepiej model radzi sobie z klasyfikacją. Możemy zauważyć, że podtyp "Basal-like" osiągnął doskonałą wartość AUC równą 1.000, co wskazuje na bardzo wysoką zdolność modelu do rozróżnienia tej klasy. Podtypy "Her2-enriched", "Luminal A" i "Luminal B" również uzyskały dobre wyniki AUC, co wskazuje na ich skuteczną klasyfikację, choć nieco niższą niż w przypadku "Basal-like".

7 Podsumowanie - zastosowanie w medycynie

Udało nam się osiągnąc dobre wyniki. Dzięki osiągnięciom w klasyfikacji raka na podstawie ekspresji białek, zbliżamy się coraz bardziej do rozwiązania, które może przyczynić się do postępu w medycynie. Zauważamy, że nasz model, pomimo mniejszej ilości dostępnych danych, radzi sobie bardzo dobrze i osiąga zadowalające wyniki. W przyszłości możemy rozważyć rozbudowę modelu, aby lepiej przewidywał podtypy, takie jak Luminal B i Her2-enriched, które do tej pory udało nam się przewidywać z dużą skutecznością, lecz zmniejszą niż w przypadku dwóch pozostałych podtypów.

Ludzkie sukcesy w tej dziedzinie mają potencjał do przekładania się na praktykę kliniczną, gdzie dokładna klasyfikacja podtypów raka może prowadzić do lepszych strategii leczenia i wyników terapeutycznych. Kontynuacja badań i rozwójów modelów pozwolą jeszcze bardziej zgłębić tę problematykę i w pełni wykorzystać potencjał proteomiki w diagnostyce i leczeniu nowotworów

Literatura

[1] Płodzich, A. *Proteomika i jej zastosowanie w wybranych jednostkach chorobowych.* Pracownia Zapewnienia Jakości, Zakład Transfuzjologii, Instytut Hematologii i Transfuzjologii.

[2] Wikipedia.

[3] National Library of Medicine.