Selekcja cech z wykorzystaniem Mutual Info Classif, RFE, F1-score Method, MRMR do budowana modeli matematycznych opartych o metody J48, SVM, 5NN

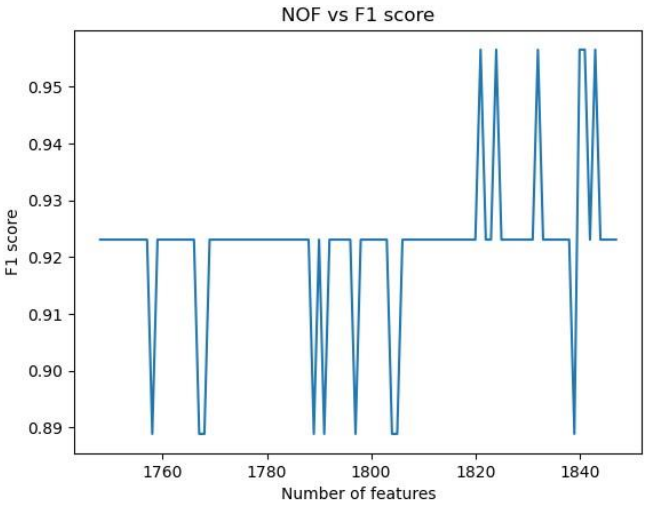
Wyliczenia przedstawione w raporcie zostały wykonane przy pomocy bibliotek języka Python. Wybrane metody selekcji niekiedy pochłaniają znaczną ilość mocy obliczeniowej oraz czasu (wymóg przeliczenia zbioru cech każdorazowo w momencie zmiany ilości pożądanych cech), przez co badania zostały ograniczone do pewnego stopnia bazując na wynikach obliczeń wcześniejszych.

F1-score Method

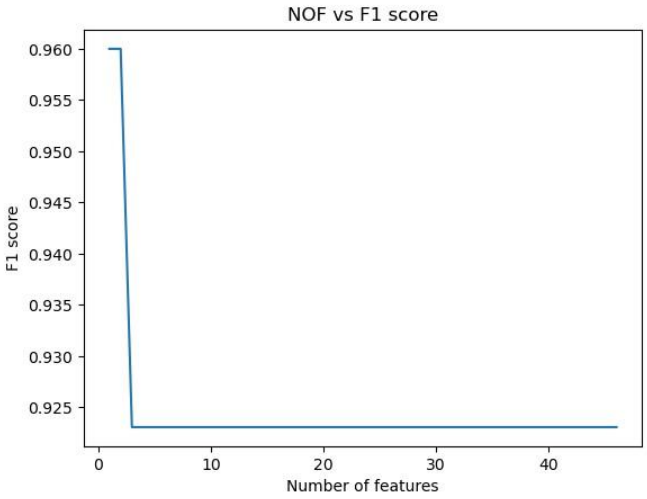
Wybrana metoda opiera się na odrzucaniu kolejnych cech bazując na wewnętrznym parametrze 'feature importance' modelu zbudowanego w oparciu o drzewo decyzyjne. Sposób obliczeń nie ma zastosowania w pozostałych modelach w związku z brakiem wspomnianego wewnętrznego parametru, który jest kluczowy przy obliczeniach.

W każdym kroku obliczeń budujemy model bazując na zbiorze 'dlbcl' lub 'prostate testing'. Następnie wyliczmy parametr f1 oraz usuwamy najmniej znaczącą cechę. Czynności powtarzamy do momentu zredukowania zbioru trenującego do wyłączenie jednej cechy. Wykonanie wykresów na podstawie otrzymanych danych pozwala nam przeanalizować przebieg wartości f1 w zależności od ilości cech. Poniżej zestawiono przykładowe charakterystyki.

1. Zbiór Danych 'dlbcl'



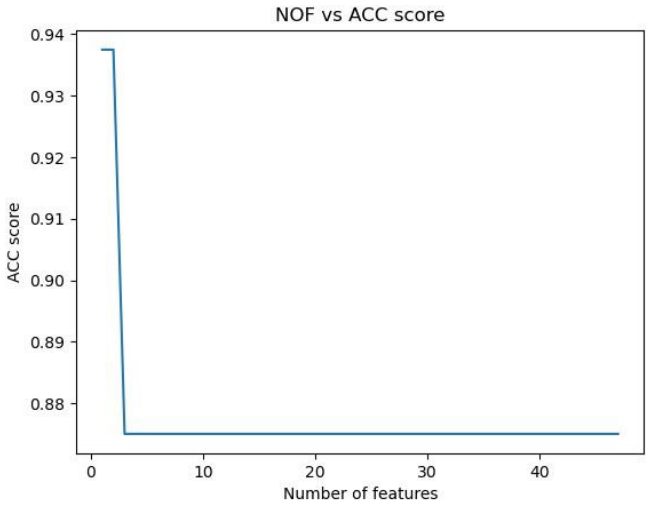
Rysunek 1 F1 score w stosunku do ilości cech w przedziale 1740-1860



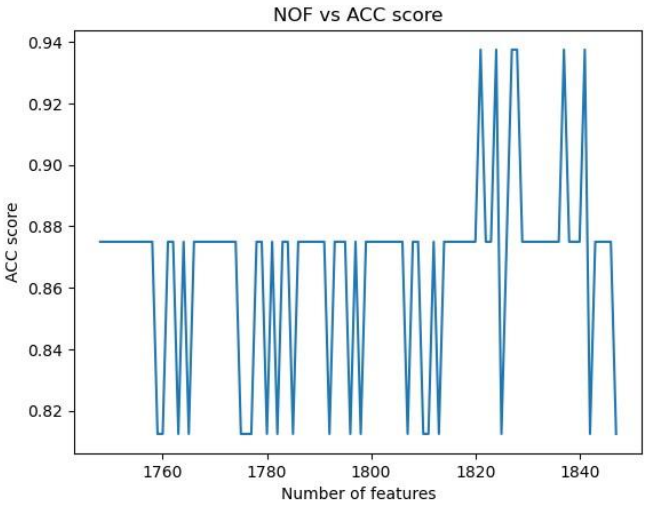
Rysunek 2 F1 score w stosunku do ilości cech w przedziale 0-60

Maksymalna wartość F1 wynosi 0.96 i nigdy nie spada poniżej 0.85. Najlepsze wyniki osiągnęły zbiory składające się odpowiednio z 1 i 2 cech ({U63743_at, Z22551_at} oraz {U63743_at}). Kolejny skok w okolice wartości maksymalnej obserwujemy w momencie przekroczenia progu 1800 cech, wcześniejsze przypadki nie osiągają wartości większej niż 0.92. Oparcie modelu o zestawy składające się 1 lub 2 cech może powodować znaczne odchylenia w przyszłości (wyniki mogą także być spowodowane ograniczonym zestawem testowym).

Zbudowane modele zostały także przeanalizowane pod kątem dokładności klasyfikacji. Wyniki pokazują zbieżność przebiegu ze wcześniejszymi charakterystykami (wartość maksymalna w początkowych dwóch przypadkach, ponowne skok po przekroczeniu progu 1800).



Rysunek 3 ACC score w stosunku do ilości cech w przedziale 0-60

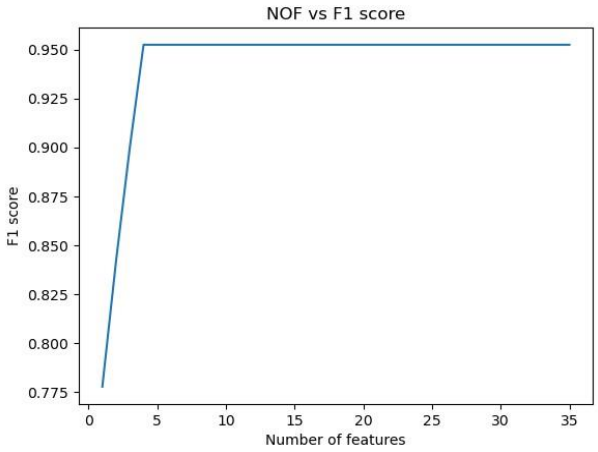


Rysunek 4 ACC score w stosunku do ilości cech w przedziale 1740-1860

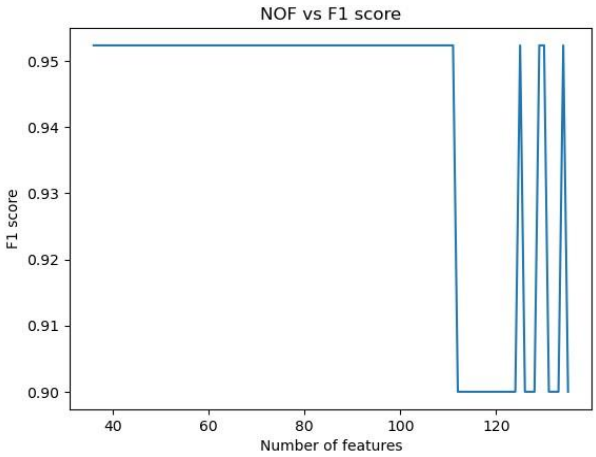
Bazując wyłącznie na dokładności klasyfikacji możemy stwierdzić, że najlepsze wyniki gwarantują zestawy złożone z 1,2 lub poszczególnych przypadków przekraczających ilość 1800 cech.

2. Zbiór danych ‘prostate testing’

Operacje wykonane na zbiorze ‘dlbcl’ zostały powtórzone na zbiorze ‘prostate testing’. Wyniki przedstawiono poniżej.

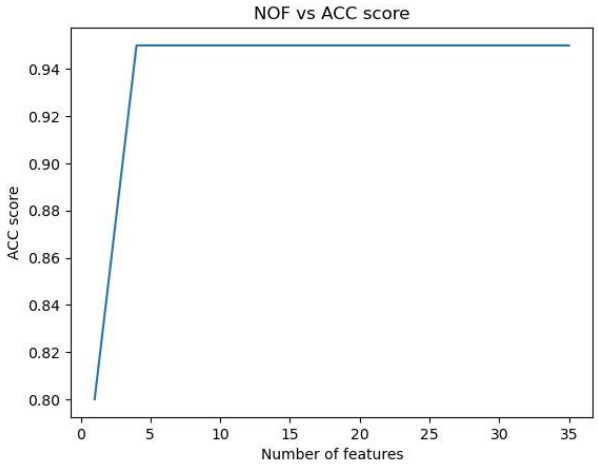


Rysunek 5 F1 score w stosunku do ilości cech w przedziale 0-35

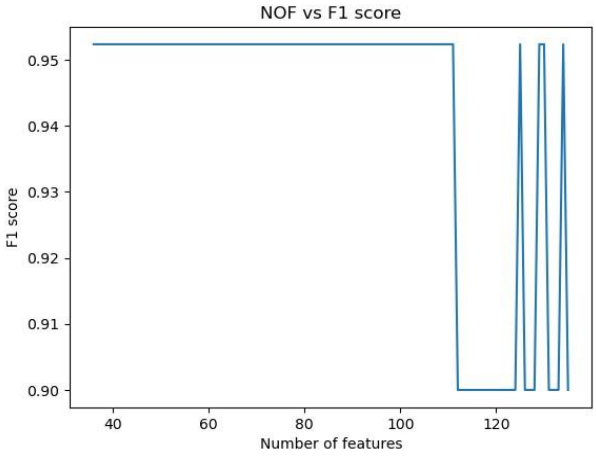


Rysunek 6 F1 score w stosunku do ilości cech w przedziale 35-140

Analizując charakterystyki zauważamy, że pierwszą wartość maksymalną osiągamy przy zbudowaniu m odelu w oparciu o zbiór 5 cech ({'41104_at', '41468_at', '41706_at', 'AFFX-M27830_5_at', 'AFFX-YEL021w /URA3_at'}). Wartość maksymalna utrzymuje się do wartości na osi X równej 111, występując następnie w momentach lokalnych gwałtownych wzrostów aż do zamknięcia przedziału cech.



Rysunek 7 ACC score w stosunku do ilości cech w przedziale 0-35



Rysunek 8 ACC score w stosunku do ilości cech w przedziale 35-140

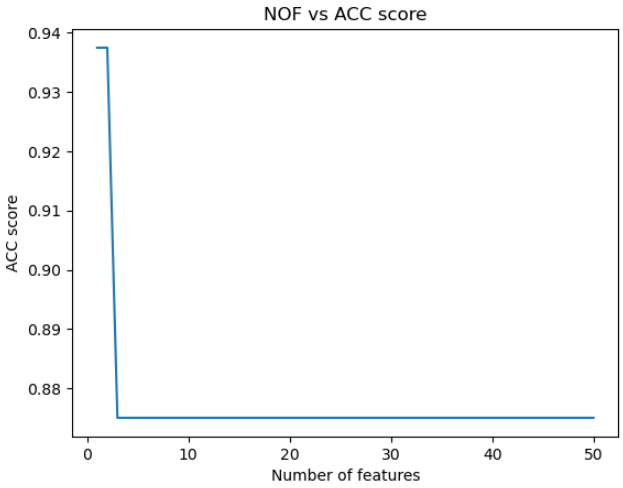
Ponownie charakterystyki przebiegu dokładności zbudowanych modeli są bardzo zbliżone od przebiegu F1 score. Bazując na otrzymanych charakterystykach optymalnym byłoby zastosowanie cech w ilości 5 – 111 wraz z późniejszymi lokalnymi maksymami. W przypadku potrzeby ograniczenia wymiarowości wybieramy zbiór z najmniejszą ilością cech. Podobnie jak w poprzednim podpunkcie ograniczony zbiór testowy może wpływać na zniekształcenie wyników.

RFE

Metoda selekcji cech RFE ze względu na większą złożoność niż poprzednia metoda wymaga znacznie większych zasobów obliczeniowych i czasowych, szczególnie w przypadku zbiorów z dużą ilością cech. Bazując na podstawie poprzednich przypadków ograniczono wyliczenia do selekcji maksymalnie 50 cech. Wszystkie przedstawione poniżej wyniki opracowano przy użyciu biblioteki pythona ‘sickit-learn’.

1. Zbiór Danych ‘dlbcl’

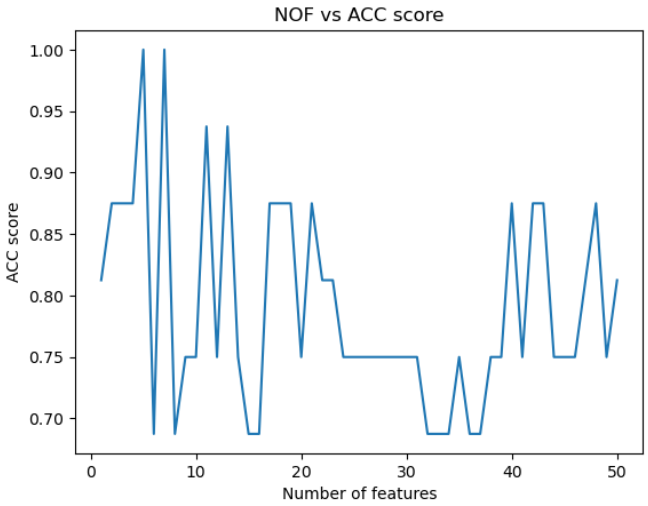
a) J48



Rysunek 9 Acc score przy budowaniu modelu metodą J48 przy pomocy 1-50 cech

Maksimum osiągamy ponownie stosując zbiory w przedziale 1-2 cech ({'U63743_at', 'U66879_at'}, {'U63743_at'}). Otrzymane wyniki nie są rozstrzygające ze względu na ich ograniczony wymiar, ale ze względu na zbieżność z wcześniejszymi wyliczeniami z zastosowaniem innych metod możemy założyć, że lokalne maksima pojawiłyby się ponownie. Otrzymane wyniki gwarantują ograniczoną wymiarowość oraz prostotę obliczeń.

b) 5NN



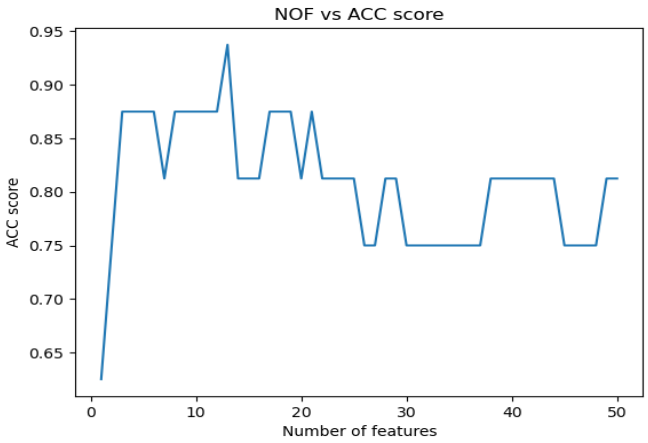
Rysunek 10 Acc score przy budowaniu modelu metodą 5NN przy pomocy 1-50 cech

Dokładność modeli budowanych w oparciu o metodę 5NN po wcześniejszej selekcji jest niestabilna w dziedzinie ilości cech już przy minimalnych wartościach, gdzie w poprzednich przypadkach przebiegi zbliżone charakterystyką pojawiały się przy znacznie większym argumencie osi X wykresu. 100% dokładność osiągamy przy selekcji 5 oraz 7 cech, odpowiednio:

- {'U63743_at', 'U66879_at', 'V00594_at', 'X95735_at', 'X95876_at' }
- {'U63743_at', 'U66879_at', 'V00594_at', 'X95632_s_at', 'X95735_at', 'X95876_at', 'X96381_rna1_at'}

Uzyskana dokładność wydaje się nierealna i prawdopodobnie spowodowana jest wąskim zbiorem testowym.

c) SVM



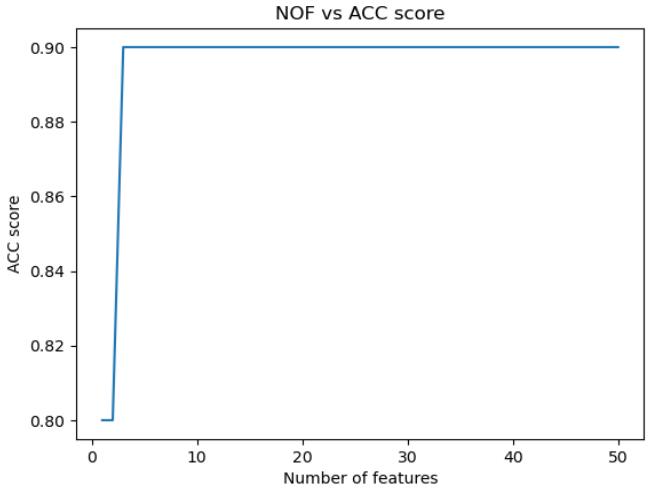
Rysunek 11 Acc score przy budowaniu modelu metodą SVM przy pomocy 1-50 cech

Charakterystyka dokładności modeli ponownie jest nieregularna i osiąga punkt maksymalny wyłącznie raz (w badanym zakresie) przy zastosowaniu zbioru z 13 cechami, odpowiednio:

- {'D28416_at','HG1980-HT2023_at', 'HG33-HT33_at', 'HG658-HT658_f_at','J04988_at', 'L17131_rna1_at','M19311_s_at', 'M63138_at', 'M63379_at', 'X02152_at', 'X03068_f_at', 'X12671_rna1_at', 'X14046_at'}

2. Zbiór danych ‘prostate testing’

a) J48

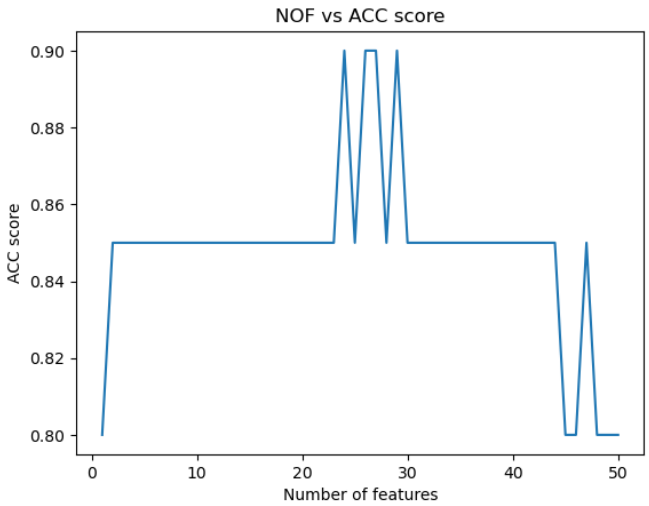


Rysunek 12 Acc score przy budowaniu modelu metodą J48 przy pomocy 1-50 cech

Dokładność modelu w badanym zakresie stabilizuje się od momentu przekroczenia progu 3 na osi X. Optymalnym, biorąc pod uwagę wyłącznie accuracy score, jest wybór podzbioru cech w liczbie 3-50. Pierwsze parę zbiorów dających wynik 90% dokładności:

- {‘41104_at', '41468_at', '41706_at'}
- {‘41104_at', '41191_at', '41468_at', '41706_at'}

b) 5NN

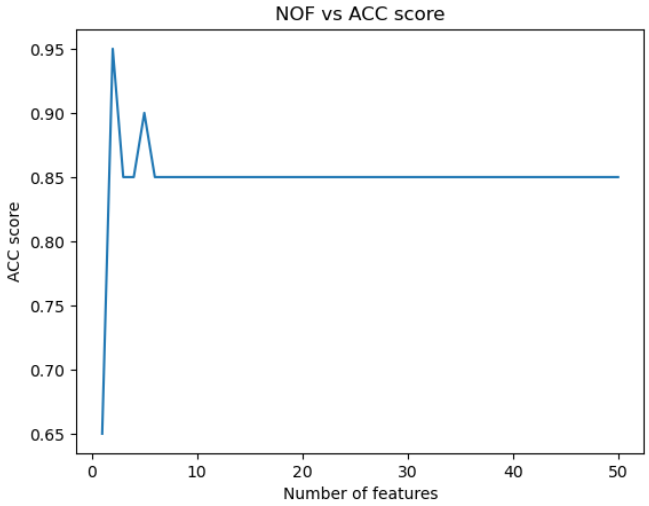


Rysunek 13 Acc score przy budowaniu modelu metodą 5NN przy pomocy 1-50 cech

W badanym zakresie osiągamy wartości maksymalne dokładności kilkukrotnie, kolejno przy zastosowaniu 23,25,26,28 cech w procesie modelowania. Wspomniane grupy cech wyglądają następująco:

- { '41104_at', '41106_at', '41111_at', '41120_at', '41123_s_at', '41134_at', '41137_at', '41138_at', '41140_at', '41143_at', '41153_f_at', '41191_at', '41468_at', '41706_at', '577_at', '585_at', '588_at', '592_at', '595_at', '607_s_at', '608_at', '609_f_at', '610_at'}
- { '41104_at', '41106_at', '41111_at', '41120_at', '41123_s_at', '41134_at', '41137_at', '41138_at', '41140_at', '41143_at', '41153_f_at', '41155_at', '41191_at', '41468_at', '41706_at', '577_at', '585_at', '588_at', '592_at', '595_at', '607_s_at', '608_at', '609_f_at', '610_at', '613_at'}
- { '41104_at', '41106_at', '41111_at', '41120_at', '41123_s_at', '41134_at', '41137_at', '41138_at', '41140_at', '41143_at', '41153_f_at', '41155_at', '41156_g_at', '41191_at', '41468_at', '41706_at', '575_s_at', '577_at', '585_at', '588_at', '592_at', '595_at', '607_s_at', '608_at', '609_f_at', '610_at'}
- { '41104_at', '41106_at', '41111_at', '41120_at', '41123_s_at', '41134_at', '41137_at', '41138_at', '41140_at', '41143_at', '41153_f_at', '41155_at', '41156_g_at', '41158_at', '41191_at', '41468_at', '41706_at', '577_at', '585_at', '588_at', '592_at', '595_at', '607_s_at', '608_at', '609_f_at', '610_at', '613_at', '614_at'}

c) SVC



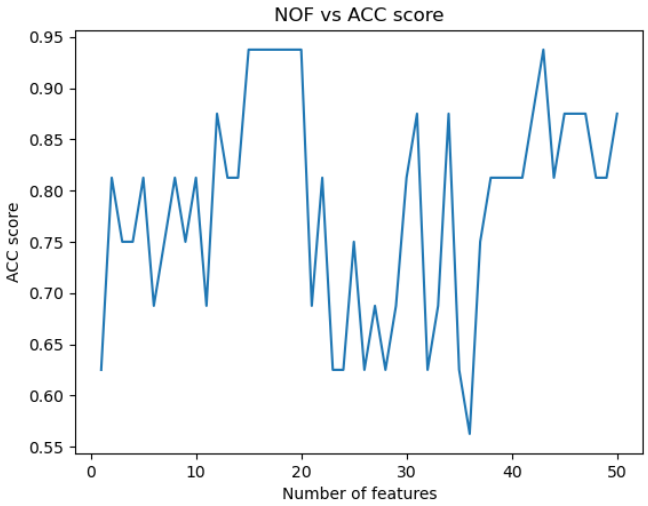
Rysunek 14 Acc score przy budowaniu modelu metodą SVM przy pomocy 1-50 cech

Dokładność modelu osiąga poziom maksymalny przy zbiorze 2 cech ({'32786_at', '40282_s_at'}), następnie się stabilizuje na poziomie 0.85.

MRMR

Podobnie jak w przypadku metody RFE MRMR pochłania znaczne zasoby obliczeniowe, czasowe, w związku z czym obliczenia zostały ograniczone do zbiorów o wielkości 50 cech. Wszystkie przedstawione poniżej wyniki zostały uzyskane przy pomocy biblioteki pythona ‘mrmr’

1. **Zbiór Danych ‘dlbcl’**
    a. *J48*

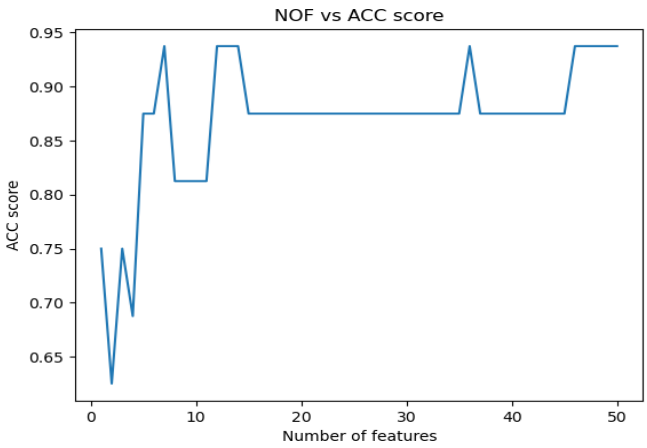


Rysunek 15 Acc score przy budowaniu modelu metodą J48 przy pomocy 1-50 cech

Przebieg dokładności modelu jest nieregularny, zauważamy wyłącznie kilka krótkich przedziałów stabilności dokładności generowanego modelu. Poziom maksymalny w badanym zakresie osiągamy z zastosowaniem kolejno 15-20, 43 cech. Kilka wybranych grup wspomnianych cech prezentuje się następująco:

- { 'Z11793_at', 'D21090_at', 'X76942_s_at', 'U50527_s_at', 'V00594_s_at', 'S73591_at', 'HG4272-HT4542_at', 'AB002409_at', 'U63743_at', 'U56814_at', 'D42041_at', 'X02152_at', 'X91911_s_at', 'M18255_cds2_s_at', 'L19686_rna1_at' }
- { 'Z11793_at', 'D21090_at', 'X76942_s_at', 'U50527_s_at', 'V00594_s_at', 'S73591_at', 'HG4272-HT4542_at', 'AB002409_at', 'U63743_at', 'U56814_at', 'D42041_at', 'X02152_at', 'X91911_s_at', 'M18255_cds2_s_at', 'L19686_rna1_at', 'U72935_cds3_s_at', 'L42324_at', 'M14328_s_at', 'M23323_s_at', 'X52773_at' }
- { 'Z11793_at', 'D21090_at', 'X76942_s_at', 'U50527_s_at', 'V00594_s_at', 'S73591_at', 'HG4272-HT4542_at', 'AB002409_at', 'U63743_at', 'U56814_at', 'D42041_at', 'X02152_at', 'X91911_s_at', 'M18255_cds2_s_at', 'L19686_rna1_at', 'U72935_cds3_s_at', 'L42324_at', 'M14328_s_at', 'M23323_s_at', 'X52773_at', 'M63138_at', 'U19495_s_at', 'Z50115_s_at', 'M91196_at', 'U05340_at', 'HG1980-HT2023_at', 'X03934_at', 'J02645_at', 'U14187_at', 'L17131_rna1_at', 'U49835_s_at', 'D38751_at', 'X62078_at', 'HG3484-HT3678_s_at', 'D79987_at', 'M63379_at', 'HG3945-HT4215_at', 'U37352_at', 'U73379_at', 'M94362_at', 'X00437_s_at', 'L36720_at', 'D90084_at' }

    b. *5NN*

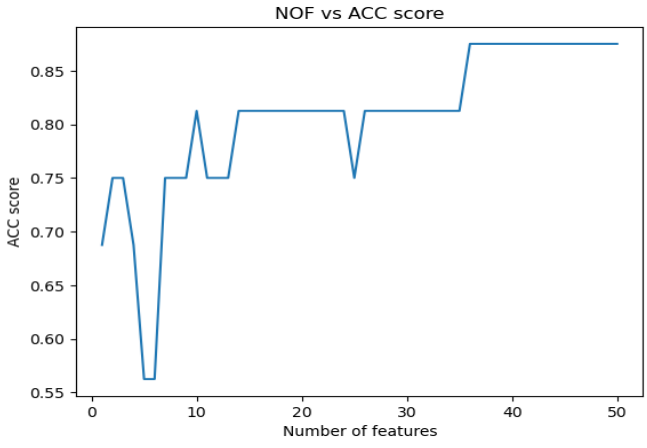


Rysunek 16 Acc score przy budowaniu modelu metodą 5NN przy pomocy 1-50 cech

Dokładność modelu stabilizuje się w pewnym zakresie. Wartość maksymalną osiągamy kilkukrotnie, zarówno przy mniejszej i znaczącej liczbie cech, kolejno 7,12,14,36,48-50. Kilka wybranych grup wspomnianych cech prezentuje się następująco:

- { 'Z11793_at', 'D21090_at', 'X76942_s_at', 'U50527_s_at', 'V00594_s_at', 'S73591_at', 'HG4272-HT4542_at'}
- { 'Z11793_at', 'D21090_at', 'X76942_s_at', 'U50527_s_at', 'V00594_s_at', 'S73591_at', 'HG4272-HT4542_at', 'AB002409_at', 'U63743_at', 'U56814_at', 'D42041_at', 'X02152_at', 'X91911_s_at', 'M18255_cds2_s_at'}
- { 'Z11793_at', 'D21090_at', 'X76942_s_at', 'U50527_s_at', 'V00594_s_at', 'S73591_at', 'HG4272-HT4542_at', 'AB002409_at', 'U63743_at', 'U56814_at', 'D42041_at', 'X02152_at', 'X91911_s_at', 'M18255_cds2_s_at', 'L19686_rna1_at', 'U72935_cds3_s_at', 'L42324_at', 'M14328_s_at', 'M23323_s_at', 'X52773_at', 'M63138_at', 'U19495_s_at', 'Z50115_s_at', 'M91196_at', 'U05340_at', 'HG1980-HT2023_at', 'X03934_at', 'J02645_at', 'U14187_at', 'L17131_rna1_at', 'U49835_s_at', 'D38751_at', 'X62078_at', 'HG3484-HT3678_s_at', 'D79987_at', 'M63379_at', 'HG3945-HT4215_at', 'U37352_at', 'U73379_at', 'M94362_at', 'X00437_s_at', 'L36720_at', 'D90084_at', 'X03350_at', 'V00594_at', 'D14662_at', 'X99076_rna1_at', 'M57710_at'}

c. SVM



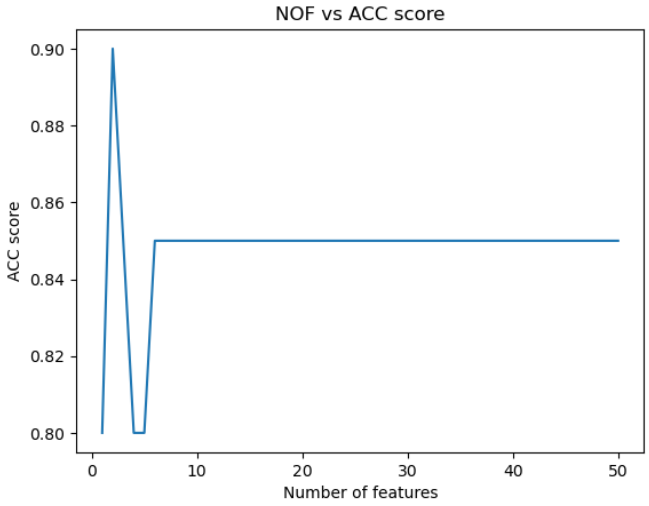
Rysunek 17 Acc score przy budowaniu modelu metodą SVM przy pomocy 1-50 cech

Dokładność modelu przez większą część dziedziny utrzymuje charakter monotonicznie narastający/stały. Wartość maksymalną osiągamy dopiero w momencie przekroczenia progu 36 cech. Parę zbiorów wspomnianych cech przedstawiono poniżej:

- {'Z11793_at', 'D21090_at', 'X76942_s_at', 'U50527_s_at', 'V00594_s_at', 'S73591_at', 'HG4272-HT4542_at', 'AB002409_at', 'U63743_at', 'U56814_at', 'D42041_at', 'X02152_at', 'X91911_s_at', 'M18255_cds2_s_at', 'L19686_rna1_at', 'U72935_cds3_s_at', 'L42324_at', 'M14328_s_at', 'M23323_s_at', 'X52773_at', 'M63138_at', 'U19495_s_at', 'Z50115_s_at', 'M91196_at', 'U05340_at', 'HG1980-HT2023_at', 'X03934_at', 'J02645_at', 'U14187_at', 'L17131_rna1_at', 'U49835_s_at', 'D38751_at', 'X62078_at', 'HG3484-HT3678_s_at', 'D79987_at', 'M63379_at'}
- {'Z11793_at', 'D21090_at', 'X76942_s_at', 'U50527_s_at', 'V00594_s_at', 'S73591_at', 'HG4272-HT4542_at', 'AB002409_at', 'U63743_at', 'U56814_at', 'D42041_at', 'X02152_at', 'X91911_s_at', 'M18255_cds2_s_at', 'L19686_rna1_at', 'U72935_cds3_s_at', 'L42324_at', 'M14328_s_at', 'M23323_s_at', 'X52773_at', 'M63138_at', 'U19495_s_at', 'Z50115_s_at', 'M91196_at', 'U05340_at', 'HG1980-HT2023_at', 'X03934_at', 'J02645_at', 'U14187_at', 'L17131_rna1_at', 'U49835_s_at', 'D38751_at', 'X62078_at', 'HG3484-HT3678_s_at', 'D79987_at', 'M63379_at', 'HG3945-HT4215_at', 'U37352_at', 'U73379_at', 'M94362_at', 'X00437_s_at', 'L36720_at', 'D90084_at', 'X03350_at', 'V00594_at', 'D14662_at'}

2. Zbiór danych ‘prostate testing’

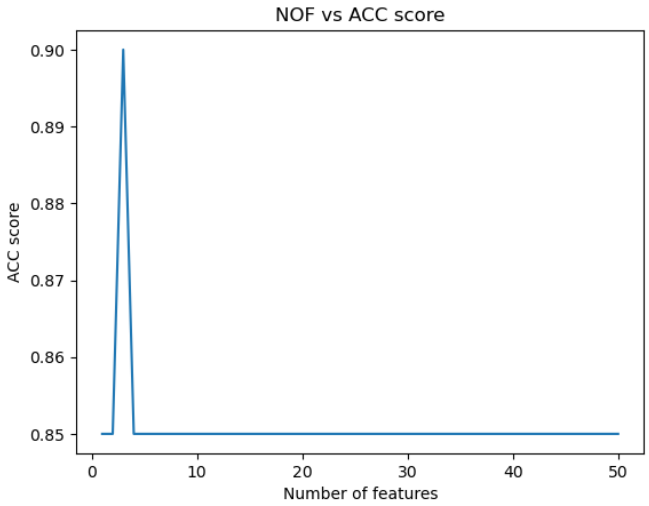
a. J48



Rysunek 18 Acc score przy budowaniu modelu metodą J48 przy pomocy 1-50 cech

Wartość maksymalną osiągamy już przy zastosowaniu zbioru 2 wybranych cech ({'37639_at', '33396_at’}). Wartości w dalszej części pomiarów stabilizują się poniżej poziomu maksymalnego.

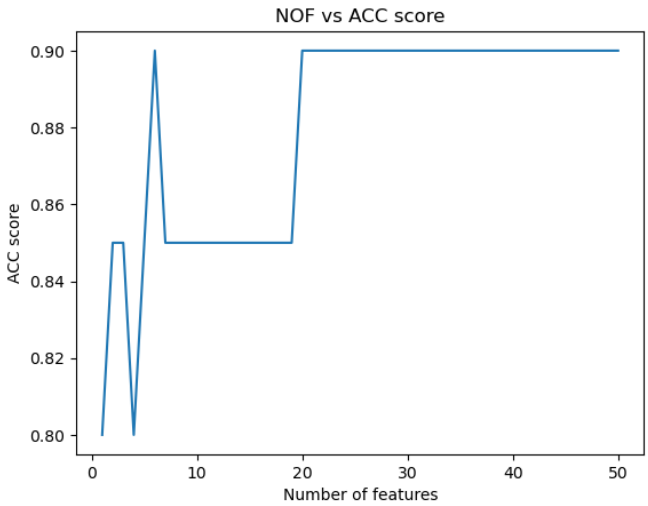
b. 5NN



Rysunek 19 Acc score przy budowaniu modelu metodą 5NN przy pomocy 1-50 cech

Dokładność modelu osiąga wartość maksymalną w badanym zakresie już przy zastosowaniu zbioru 3 cech ({'37639_at', '33396_at', '32275_at' }). Maksimum jest jedynym punktem odchylenia od stabilności charakterystyki.

c. SVM



Rysunek 20 Acc score przy budowaniu modelu metodą SVM przy pomocy 1-50 cech

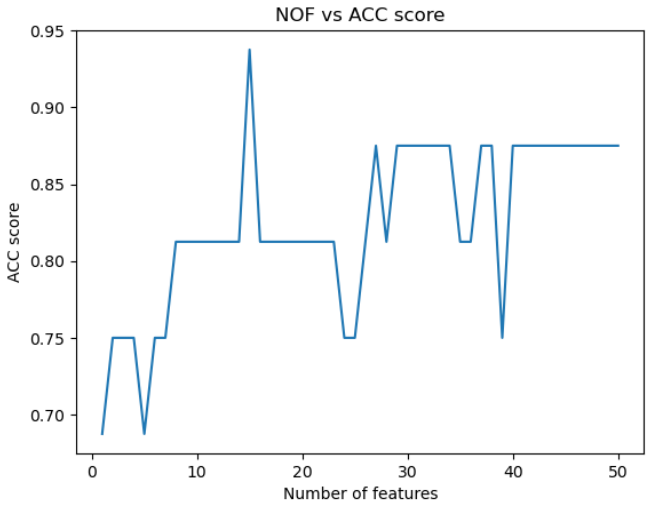
Charakterystyka przez większą część dziedziny utrzymuje charakter stabilny. Miejscami pojawiają się lokalne wzrosty, spadki dla wybranych zbiorów cech. Wartości maksymalne osiągamy kilkukrotnie w badanym przedziale z zastosowaniem 17, 20-50 cech. Poniżej parę wspomnianych zbiorów:

- { '37639_at', '33396_at', '32275_at', '41706_at', '41468_at', '32598_at', '37366_at', '575_s_at', '38634_at', '34840_at', '36491_at', '35276_at', '40282_s_at', '36174_at', '38057_at', '38827_at', '31791_at' }
- { '37639_at', '33396_at', '32275_at', '41706_at', '41468_at', '32598_at', '37366_at', '575_s_at', '38634_at', '34840_at', '36491_at', '35276_at', '40282_s_at', '36174_at', '38057_at', '38827_at', '31791_at', '1740_g_at', '38291_at', '39756_g_at', '38059_g_at', '556_s_at', '39073_at', '38087_s_at', '41732_at', '34775_at', '32109_at', '38469_at', '914_g_at', '39054_at', '37000_at', '1251_g_at', '38028_at', '41696_at', '35178_at', '37599_at', '40071_at', '32485_at' }

Mutual Info Classif

Podobnie jak w przypadku poprzednich metod ‘Mutual Info Classif’ pochłania znaczne zasoby obliczeniowe, czasowe, w związku z czym obliczenia zostały ograniczone do zbiorów o wielkości 50 cech. Wszystkie przedstawione poniżej wyniki zostały uzyskane przy pomocy biblioteki pythona ‘sklearn-relief’

1. **Zbiór danych ‘dlbcl’**
   a. J48

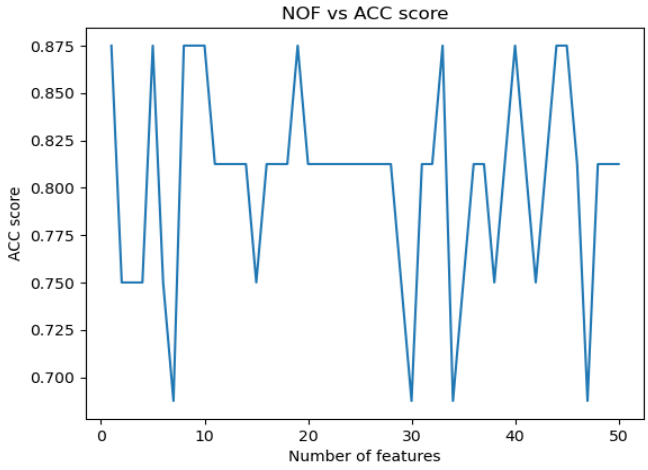


Rysunek 21 Acc score przy budowaniu modelu metodą J48 przy pomocy 1-50 cech

Dokładność modelu osiąga jednorazowo wartość maksymalną i dynamicznie się zmienia. Największy wynik osiągnięto przy zastosowaniu zbioru 15 cech:

- {‘D79987_at' 'D79997_at' 'D86973_at' 'HG1980-HT2023_at' 'HG4074-HT4344_at' 'J04031_at' 'L36720_at' 'M25753_at' 'U41515_at' 'U63743_at' 'X02152_at' 'X52142_at' 'X62078_at' 'X65550_at' 'X69433_at’}

b. 5NN

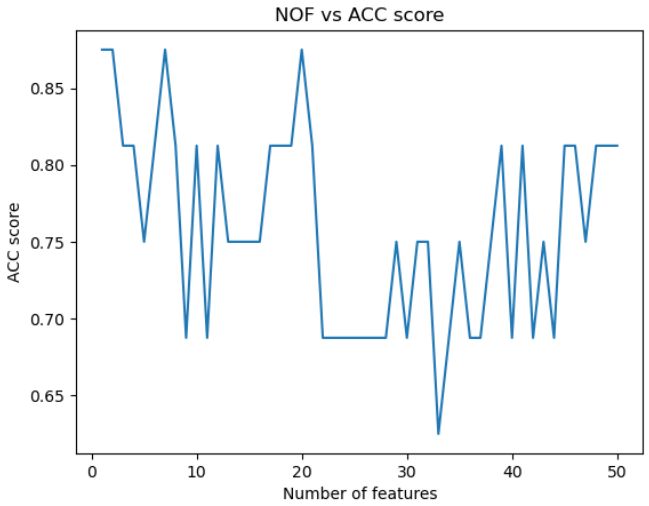


Rysunek 22 Acc score przy budowaniu modelu metodą 5NN przy pomocy 1-50 cech

Dokładność modelu zmienia się z dużą dynamiką kilkukrotnie osiągając wartość maksymalną. Najlepsze wyniki zapewniają kolejno zbiory złożone z 5,8,9,10,33,44-45. Poniżej zestawiono parę wspomnianych zbiorów:

- {‘D79997_at' 'D86973_at' 'J04031_at' 'V00594_at' 'X62078_at' }
- {‘D79997_at' 'D86973_at' 'J04031_at' 'L36720_at' 'M25753_at' 'U41515_at' 'U63743_at' 'V00594_at' 'X62078_at’}

c. SVC



Rysunek 23 Acc score przy budowaniu modelu metodą SVC przy pomocy 1-50 cech

Charakterystyka ponownie utrzymuje dynamicznie charakter miejscowo osiągając wartość maksymalną. Najlepsze wyniki zapewniają kolejno zbiory złożone z 1,2,7,20 cech. Poniżej zestawiono parę wspomnianych zbiorów:

- {‘J04031_at’}
- {‘D79997_at' 'D86973_at' 'HG1980-HT2023_at' 'J04031_at' 'L19686_rna1_at’ ‘X02152_at' 'X62078_at' }