PRACOWNIA ANALIZY DANYCH II

PROJEKT I

BINARNA KLASYFIKACJA TYPÓW WINA Z WYKORZYSTANIEM MODELI REGRESJI

MAŁGORZATA RYLL, MARCEL GOŁĄB

1. Cel projektu

Naszym zadaniem jest porównanie skuteczności trzech różnych modeli regresyjnych w klasyfikacji binarnej danych dotyczących win. W ramach projektu zastosujemy następujące podejścia:

1. **Regresja logistyczna** – klasyczna metoda probabilistyczna, która umożliwia klasyfikację binarną na podstawie dostępnych cech charakteryzujących wino.
2. **Regresja logistyczna z *PCA*** – podejście polegające na redukcji wymiarowości danych za pomocą analizy głównych składowych (*PCA*), a następnie zastosowaniu regresji logistycznej do klasyfikacji.
3. **Relaxed LASSO** – metoda regresyjna integrująca selekcję zmiennych i regularyzację, co pozwala na wskazanie kluczowych cech odróżniających wina białe od czerwonych.

Celem projektu jest określenie, która z tych metod najefektywniej klasyfikuje wina jako białe lub czerwone, bazując na cechach chemicznych i sensorycznych.

2. Zbiór danych – *Wine Quality*

Omówienie danych

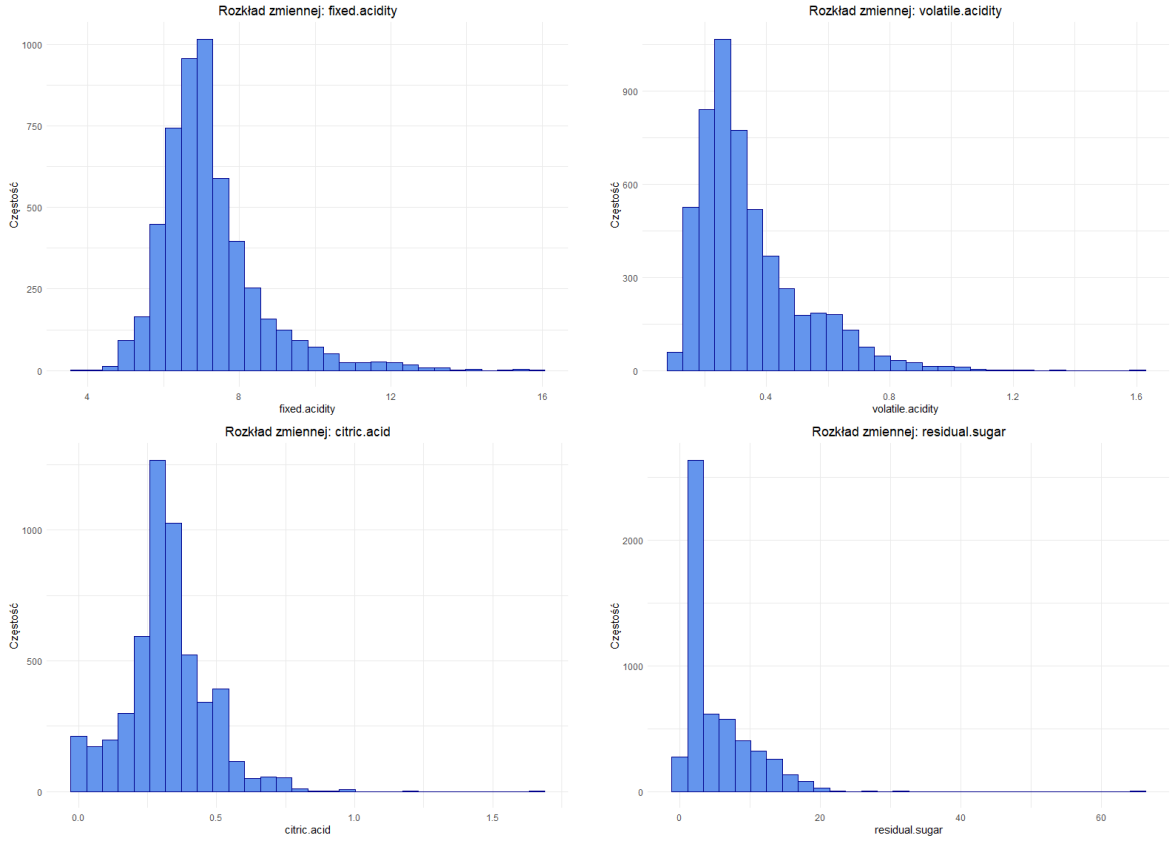
Dane, na których pracujemy, pochodzą z otwartego zbioru *Wine Quality* dostępnego na stronie *https://archive.ics.uci.edu/dataset/186/wine+quality* i dotyczą charakterystyki chemicznej oraz sensorycznej win. Zbiór danych zawiera obserwacje, które obejmują następujące atrybuty:

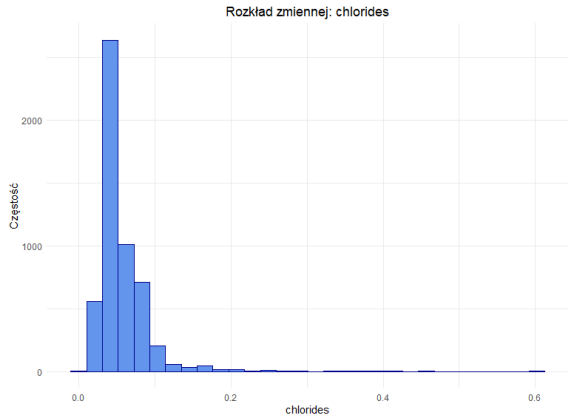
- ***Fixed Acidity***: Stała kwasowość (g/l) – kwasowość, która pozostaje po zakończeniu procesu fermentacji.
- ***Volatile Acidity***: Lotna kwasowość (g/l) – związki kwasowe, które mogą być wykrywalne jako zapach octu.
- ***Citric Acid***: Kwas cytrynowy (g/l) – wpływa na świeżość i równowagę smaku wina.
- ***Residual Sugar***: Cukier resztkowy (g/l) – ilość cukru, która nie została przekształcona w alkohol podczas fermentacji.
- ***Chlorides***: Zawartość chlorków (g/l) – wskaźnik zasolenia wina.
- ***Free Sulfur Dioxide***: Wolny dwutlenek siarki (mg/l) – forma SO_2, która chroni wino przed utlenianiem i rozwojem mikroorganizmów.
- ***Total Sulfur Dioxide***: Całkowity dwutlenek siarki (mg/l) – łączna zawartość SO_2, obejmująca wolną i związaną formę.
- ***Density***: Gęstość (g/cm^3) – wskaźnik ilości cukrów i alkoholu w winie.

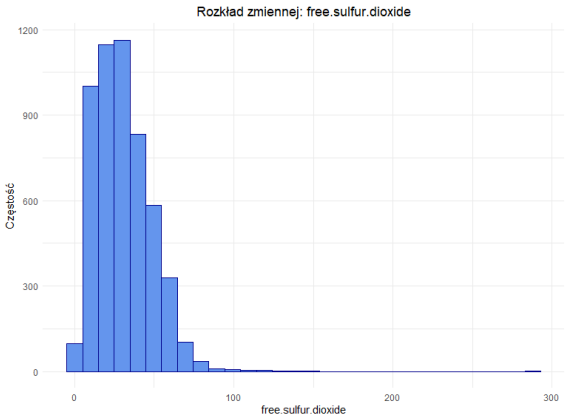
- ***pH***: Wskaźnik kwasowości – miara, jak bardzo kwasowe jest wino.
- ***Sulphates***: Zawartość siarczanów (g/l) – wpływają na smak oraz działają jako naturalny konserwant.
- ***Alcohol***: Zawartość alkoholu (%) – kluczowy czynnik wpływający na jakość wina.
- ***Quality***: Ocena jakości wina (skala od 0 do 10) – sensoryczna ocena dokonana przez ekspertów.
- ***Type***: Typ wina (zmienna celu) – wskazuje, czy wino jest białe, czy czerwone:
    - ***0***: Wino białe
    - ***1***: Wino czerwone

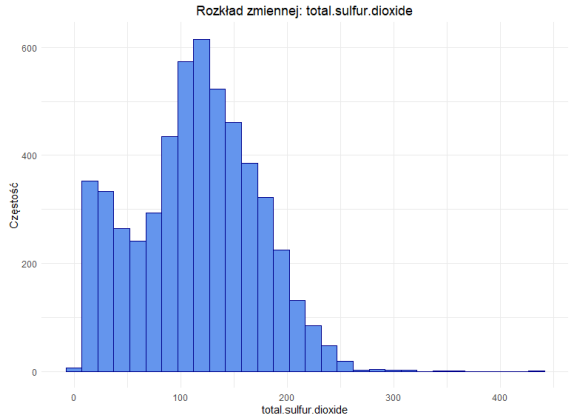
Rozkłady zmiennych

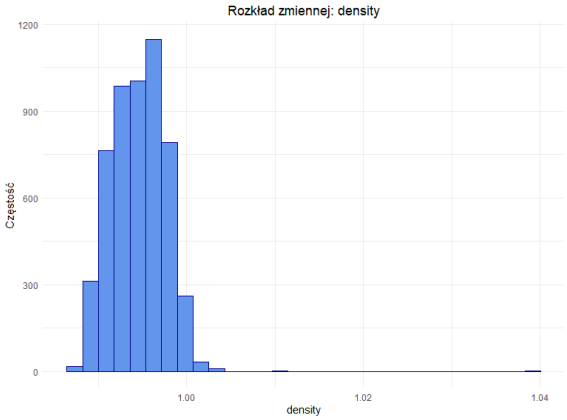
W celu analizy rozkładu danych oraz zrozumienia wartości zmiennych, wizualizujemy dane za pomocą histogramów. Dla zmiennych *quality* i *type* wykorzystujemy wykresy kolumnowe, które zapewniają lepszą czytelność i ułatwiają interpretację wyników.

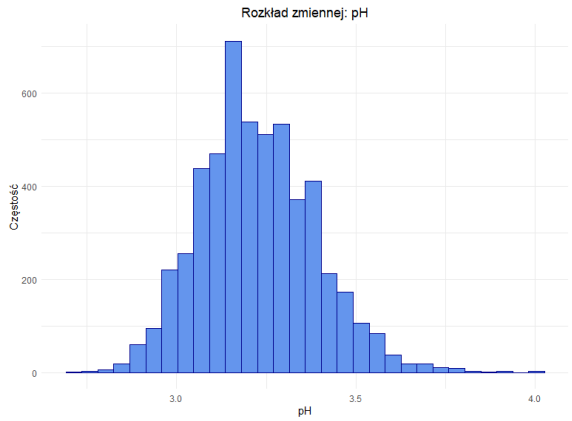


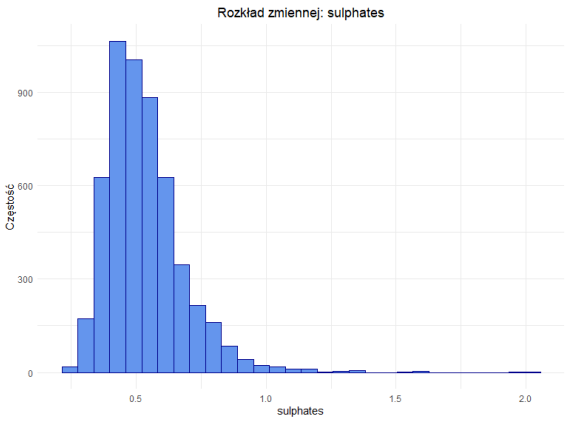

Rozkład zmiennej: chlorides
Częstość
chlorides

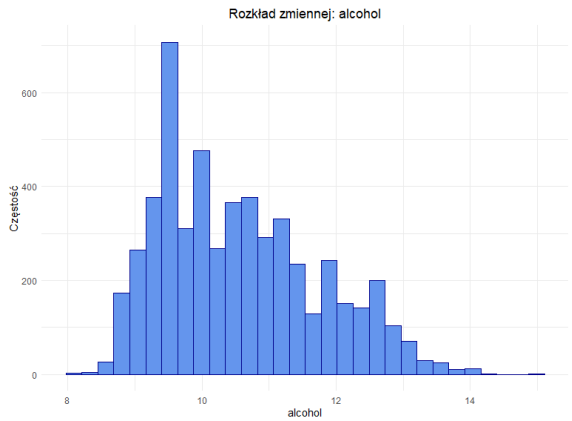

Rozkład zmiennej: free.sulfur.dioxide
Częstość
free.sulfur.dioxide

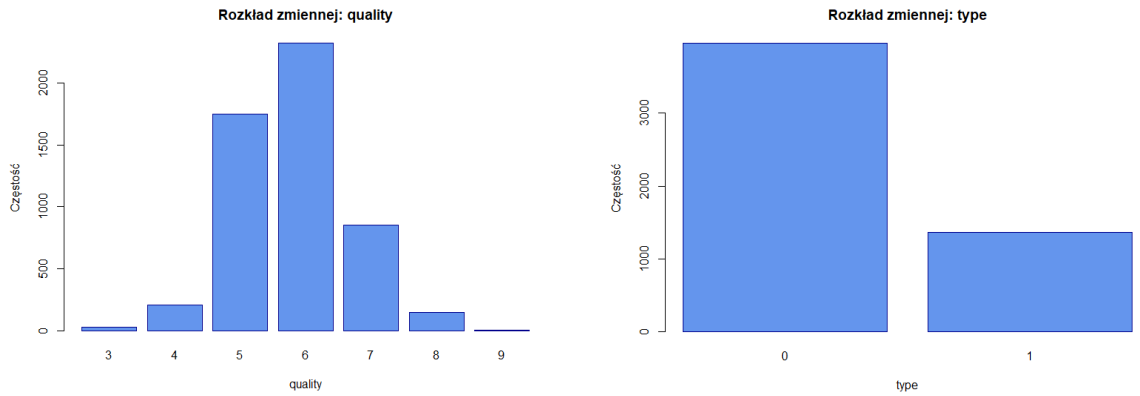

Rozkład zmiennej: total.sulfur.dioxide
Częstość
total.sulfur.dioxide


Rozkład zmiennej: density
Częstość
density


Rozkład zmiennej: pH
Częstość
pH


Rozkład zmiennej: sulphates
Częstość
sulphates


Rozkład zmiennej: alcohol
Częstość
alcohol

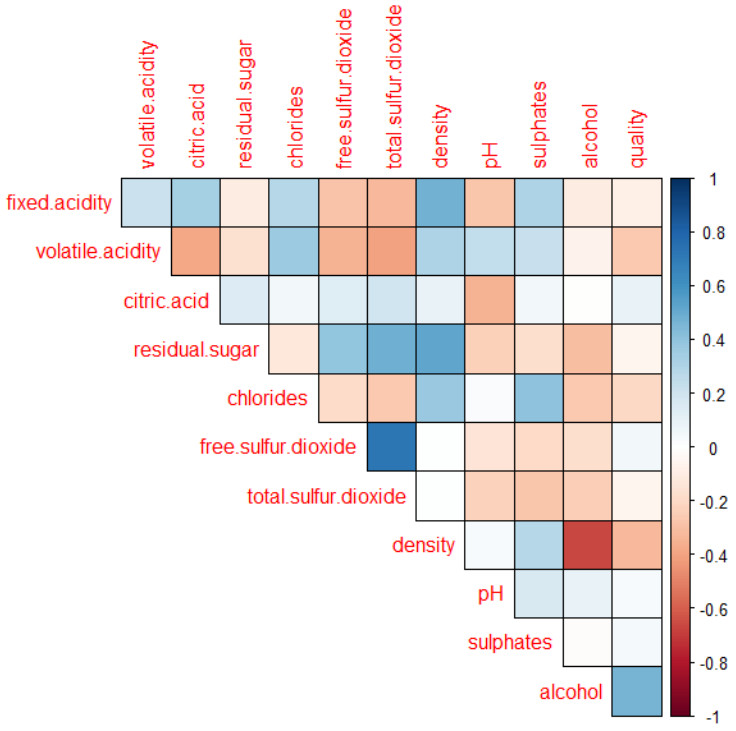


Dane zawierają 1359 unikalnych obserwacji dotyczących win czerwonych oraz 3961 dotyczących win białych. Dysproporcja jest zauważalna i istnieje ryzyko, że modele będą częściej przewidywać klasę win białych, ale nie integrujemy w ilość obserwacji. Modele będą musiały sobie z tą dysproporcją poradzić.

Zauważamy obecność wartości odstających. Po ich przeanalizowaniu dochodzimy jednak do wniosku, że różnice te nie są na tyle znaczące, aby jednoznacznie uznać je za błędne pomiary lub wyniki odbiegające od rzeczywistości. W związku z tym przyjmujemy, że mogą one reprezentować naturalne zróżnicowanie w badanej próbie i zostały uwzględnione w dalszych etapach analizy.

Korelacja między zmiennymi

Macierz korelacji



Wszystkie dane objaśniające są zmiennymi numerycznymi. Analiza korelacji pozwala zaobserwować kilka istotnych zależności:

- **Wysoka dodatnia korelacja**:
  - Pomiędzy *free.sulfur.dioxide* a *total.sulfur.dioxide*, co jest logiczne, ponieważ całkowity dwutlenek siarki obejmuje również jego wolną formę.
- **Negatywna korelacja**:
  - Pomiędzy *density* a *alcohol*. Im wyższa zawartość alkoholu, tym niższa gęstość wina, co również jest zgodne z właściwościami fizycznymi wina, gdyż alkohol rozrzedza ciecz.
- **Brak silnych korelacji**:
  - Większość zmiennych ma współczynniki korelacji poniżej |0.6|, co wskazuje na ich względną niezależność.

Ogólnie dane są odpowiednie do dalszej analizy, a zastosowanie metod takich jak *PCA* lub regularyzacja może pomóc w radzeniu sobie z potencjalną współliniowością.

Podział danych

Dane zostały podzielone na zbiór treningowy i testowy, gdzie 70% obserwacji przeznaczono na trenowanie modelu, a pozostałe 30% na jego testowanie. Taki podział jest standardową praktyką w modelowaniu predykcyjnym, ponieważ pozwala na efektywne dopasowanie modelu do danych, a jednocześnie umożliwia jego ocenę na niezależnym zbiorze danych, co zwiększa wiarygodność wyników.

Do przeprowadzenia podziału wykorzystano funkcję *createDataPartition* z pakietu *caret*, która zapewniła zrównoważony rozkład zmiennej docelowej (*type*) w obu zbiorach. Dzięki temu zarówno zbiór treningowy, jak i testowy odzwierciedlają proporcje klas win białych i czerwonych w całym zbiorze danych.

3. Modele regresji

3.1. Model regresji logistycznej

Opis modelu

Budując model regresji logistycznej chcemy uzyskać kompromis między dopasowaniem modelu a jego złożonością, dlatego dokonujemy procesu selekcji zmiennych za pomocą kryterium *Akaike'a*.

Selekcja zmiennych

Proces selekcji zmiennych przeprowadzono za pomocą funkcji *stepAIC* z metodą eliminacji wstecznej (direction = "backward"), rozpoczynając od modelu zawierającego wszystkie dostępne zmienne objaśniające.

Wartość AIC modelu początkowego wynosi 261.86. Po zakończeniu selekcji jedynymi wykluczonymi zmiennymi zostały *citric.aid* oraz *sulphates*, co dało nam model końcowy składający się z 10 predyktorów. I wartością AIC równą 259.74.

Chociaż różnica 2.12 jednostek jest stosunkowo niewielka, uzyskaliśmy uproszczony model, co jest korzystniejsze w kontekście interpretacji oraz oszczędności obliczeniowej.

Wnioski z modelu

```
Call:
glm(formula = type ~ fixed.acidity + volatile.acidity + residual.sugar +
    chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
    density + pH + alcohol + quality, family = "binomial", data = train_data)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.585e+03  2.483e+02 -10.412  < 2e-16 ***
fixed.acidity        -1.010e+00  2.863e-01  -3.529 0.000417 ***
volatile.acidity      6.244e+00  1.200e+00   5.202 1.97e-07 ***
residual.sugar       -9.121e-01  1.087e-01  -8.394  < 2e-16 ***
chlorides             1.809e+01  4.799e+00   3.769 0.000164 ***
free.sulfur.dioxide   4.775e-02  1.767e-02   2.703 0.006873 **
total.sulfur.dioxide -5.391e-02  6.172e-03  -8.734  < 2e-16 ***
density               2.588e+03  2.521e+02  10.269  < 2e-16 ***
pH                   -3.314e+00  1.797e+00  -1.845 0.065094 .
alcohol               2.672e+00  3.709e-01   7.205 5.79e-13 ***
quality               5.510e-01  2.691e-01   2.048 0.040603 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4234.38  on 3724  degrees of freedom
Residual deviance:  237.74  on 3714  degrees of freedom
AIC: 259.74

Number of Fisher Scoring iterations: 9
```

Na podstawie podanych wyników modelu regresji logistycznej możemy wyróżnić kilka kluczowych obserwacji.

Znak współczynników:

Sześć na jedenaście współczynników są dodatnie, co oznacza, że wyższe wartości tych zmiennych zwiększają log-odds wystąpienia czerwonego wina. Pozostałe współczynniki są ujemne, co oznacza, że wyższe wartości tych zmiennych zmniejszają log-odds wystąpienia czerwonego wina.

Najbardziej istotna zmienna:

Zmienna *density* jest zmienną o największym wpływie na przewidywanie klasy wina, co wynika z dwóch powodów:

- Wysoka wartość współczynnika (*Estimate* = 2588), co przekłada się na to, że niewielka zmiana gęstości wina znacząco wpływa na log-odds przypisania wina do klasy czerwonej.
- P-wartość dla tej zmiennej (< 2e-16) jest bardzo niskie, co wskazuje ma bardzo wysoki poziom istotności statystycznej.

Na wcześniejszej zaprezentowanej macierzy korelacji widać silne powiązanie *density* z innymi zmiennymi.

Całość sprawia, że zmienna *density* jest silnym predyktorem różnic między winami, ponieważ czerwone wina zwykle charakteryzują się większą ilością ekstraktu, który wpływa na gęstość. Wysoki współczynnik przy tej zmiennej wskazuje na jej kluczowe znaczenie dla modelu, co jest zgodne z oczekiwaniami w kontekście cech fizycznych win.

Zmienna o najmniejszym wpływie:

Mimo że *pH* nie ma współczynnika o najmniejszej wartości bezwzględnej, jego p-wartość (0.065) wskazuje, że wpływ tej zmiennej nie jest istotny statystycznie na poziomie istotności |0.05|. Może to wynikać z faktu, że inne zmienne, takie jak *fixed.acidity* i *volatile.acidity,* lepiej opisują kwasowość wina i przejmują część tej roli w modelu.

Null Deviance oraz Residual Deviance:

Null Deviance (4234.38): Wysoka wartość *null deviance* wskazuje, że dane charakteryzują się znaczną zmiennością, której model bazowy (przewidujący jedną stałą wartość dla wszystkich obserwacji, bez uwzględnienia zmiennych objaśniających) nie jest w stanie wyjaśnić. Oznacza to, że model bazowy nie dopasowuje się do danych w sposób adekwatny.

Residual Deviance (237.74): Znaczący spadek *deviance* po uwzględnieniu zmiennych objaśniających wskazuje, że model końcowy skutecznie opisuje zmienność

odpowiedzi. Niska wartość *residual deviance* oznacza, że różnica między przewidywaniami modelu a rzeczywistymi wartościami jest niewielka, co świadczy o jego dobrym dopasowaniu.

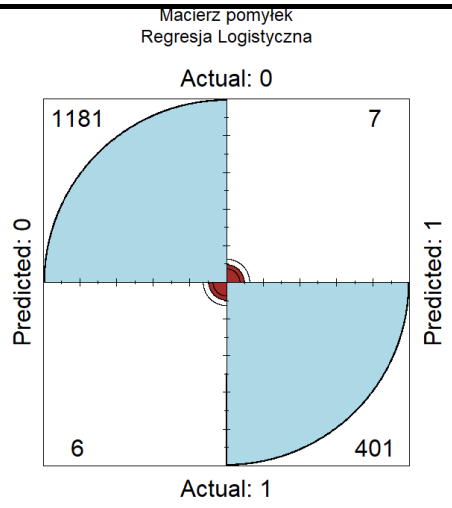
Spadek z 4234.38 do 237.74 pokazuje, że uwzględnienie zmiennych objaśniających znacząco poprawiło jakość modelu w wyjaśnianiu danych.

Iteracja Fisher Scoring:

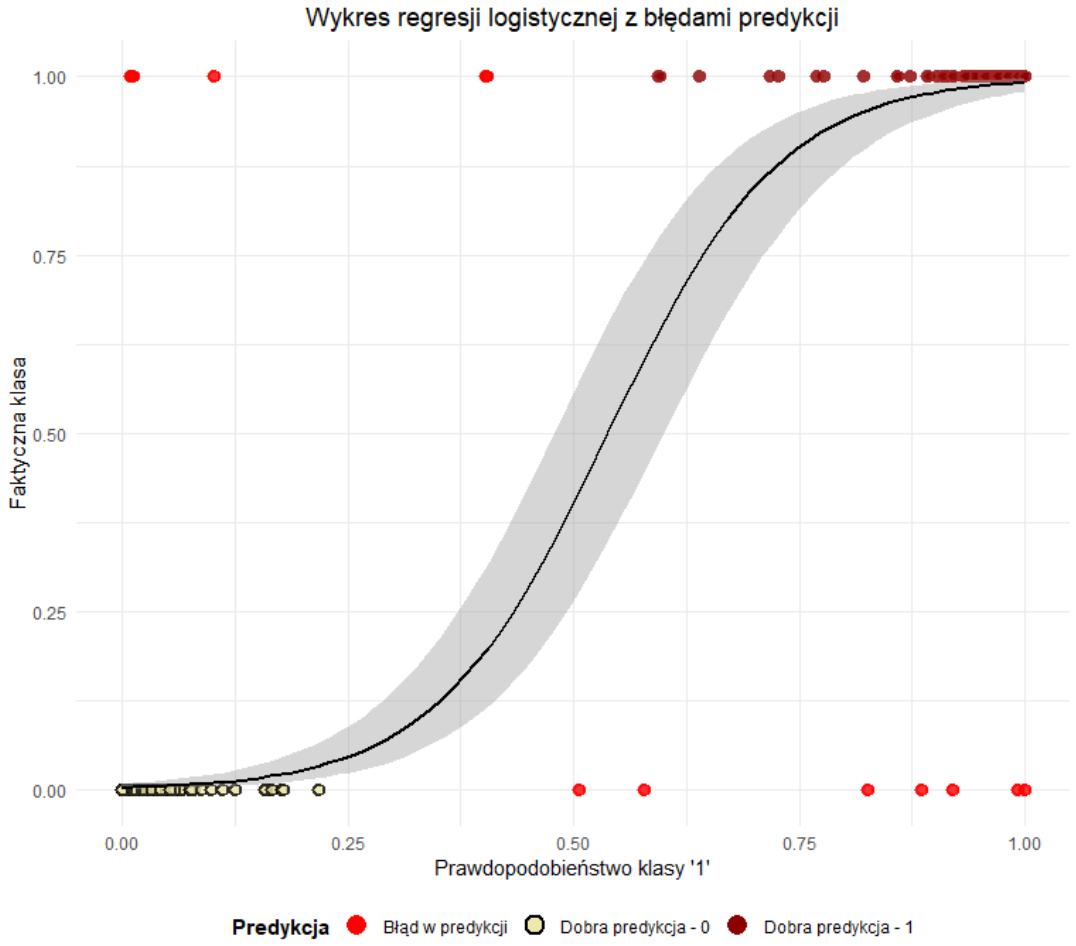
Model osiągnął zbieżność po 9 iteracjach metody Fisher Scoring. Liczba iteracji wskazuje, że model był umiarkowanie złożony, ale proces optymalizacji przebiegł stabilnie. Gdyby liczba iteracji była wyższa, mogłoby to sugerować trudności z dopasowaniem modelu do danych lub problemy z niestabilnością parametrów.

Ocena jakości modelu

Macierz pomyłek

Macierz pomyłek	Dokładność modelu
	99.18495%

Wykres krzywej sigmoidalnej



Czy model jest dobry?

Na podstawie wyników można stwierdzić, że model regresji logistycznej dobrze spełnia swoje zadanie:

1. **Wysoka dokładność na zbiorze testowym** – osiągnięcie dokładności na poziomie 99.18% wskazuje, że model poprawnie klasyfikuje zdecydowaną większość obserwacji, co świadczy o jego skuteczności.
2. **Niska liczba błędów** – jedynie 13 błędnie sklasyfikowanych obserwacji na zbiorze testowym pokazuje, że model dobrze radzi sobie z przewidywaniem typu wina.
3. **Istotność zmiennych** – większość zmiennych w modelu okazała się statystycznie istotna, co zwiększa jego wiarygodność. Szczególnie zmienna *density*, która wykazuje największy wpływ na przewidywania, jest zgodna z intuicją opartą na właściwościach fizycznych wina.
4. **Kompromis między złożonością a dopasowaniem** – zastosowanie selekcji zmiennych za pomocą kryterium *Akaike'a* pozwoliło na uzyskanie modelu o uproszczonej strukturze (10 predyktorów) przy jednoczesnym zachowaniu wysokiej jakości dopasowania.

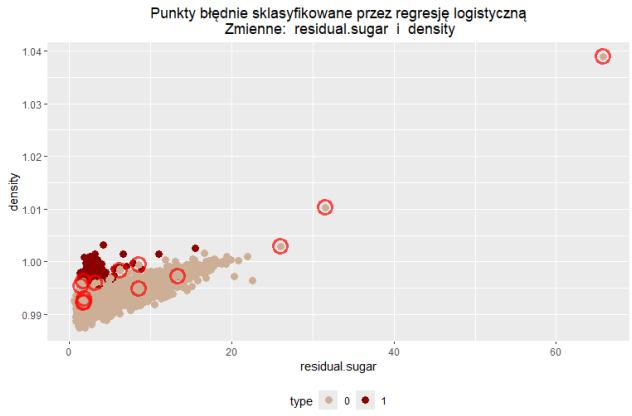
5. **Spadek *Null Deviance* i *Residual Deviance*** – duży spadek wartości *deviance* wskazuje, że model wyjaśnia znaczną część zmienności w danych, co potwierdza jego trafność.

Podsumowując, model jest dobrze dopasowany do danych, osiąga wysoką skuteczność predykcyjną i został zoptymalizowany pod kątem prostoty i interpretowalności.

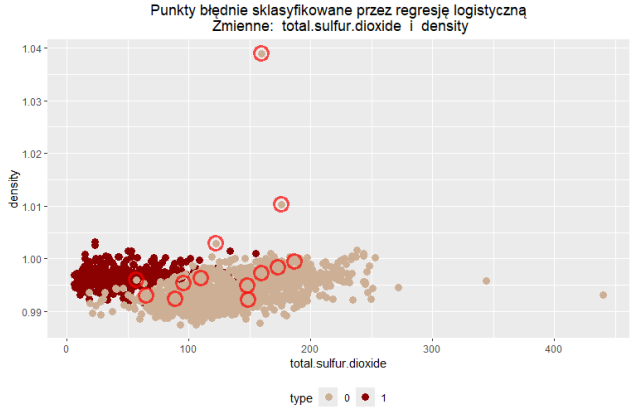
Analiza błędów

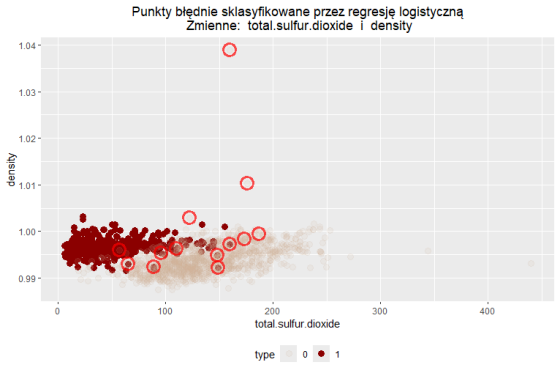
Nasza analiza opiera się na wykresach punktowych przedstawiających zależności między dwiema zmiennymi. Spośród nich pięć to standardowe wykresy punktowe, gdzie jasne punkty reprezentują wina białe, a bordowe – wina czerwone. Błędne klasyfikacje zostały wyróżnione czerwonymi okręgami. Na dwóch z tych wykresów zastosowano efekt przezroczystości, co pozwala lepiej zobrazować zagęszczenie oraz rozmieszczenie obserwacji.

Dodatkowo, jeden z wykresów to wykres typu *jitter*, na którym wprowadzono niewielki losowy szum, aby rozproszyć punkty i zapobiec ich nakładaniu się. Na tym wykresie przynależność do klasy została oznaczona kształtem punktów, a poprawność klasyfikacji – kolorem. Błędne klasyfikacje wyróżniono ciemnoczerwonym kolorem, co ułatwia ich identyfikację.

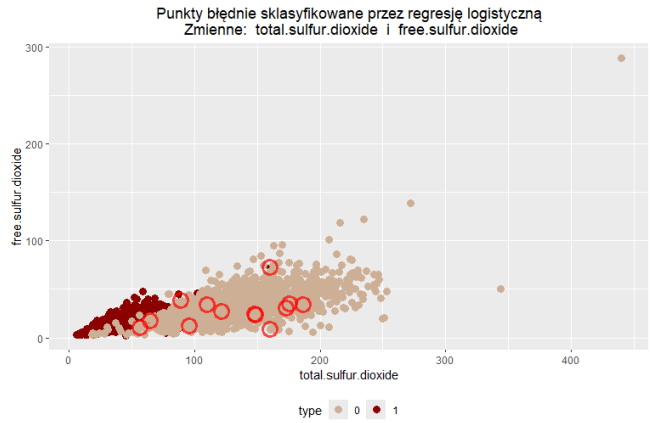


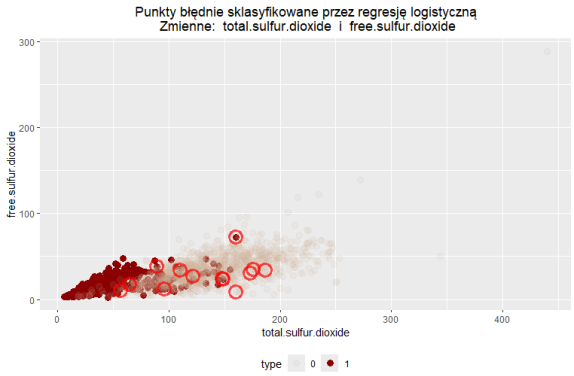
Należy zauważyć, że model popełnił trzy błędy na trzech obserwacjach, które znajdują się najdalej od pozostałych, w pobliżu grupy gęsto skonsolidowanych obserwacji. Dodatkowo, można zaobserwować, że kolory win silnie się grupują, przy czym czerwone wina mają zazwyczaj wyższą gęstość oraz bardziej skoncentrowany poziom cukru resztkowego w zakresie 0-5. W przypadku, gdy obserwacja znajduje się blisko tej granicy, model częściej popełnia błędy w klasyfikacji.





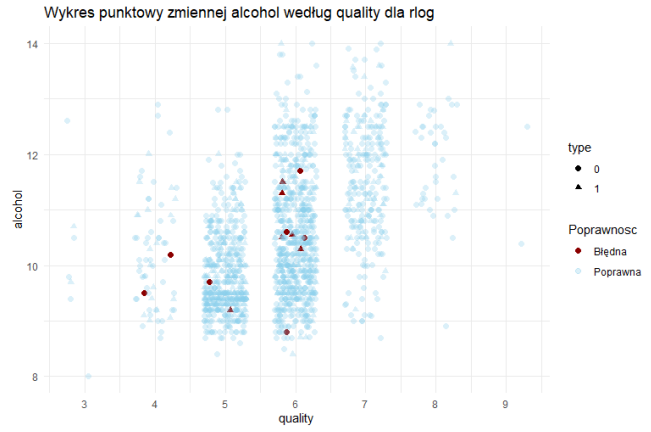
Ponownie zwracamy uwagę na punkty odstające. W przypadku zmiennej *density* model ma trudności z poprawnym przypisaniem tych obserwacji, podczas gdy dla zmiennej *total.sulfur.dioxide* radzi sobie znacznie lepiej. Na tym wykresie zastosowano zmienioną przezroczystość kolorów, co pozwala lepiej dostrzec zagęszczenie obserwacji. Dzięki tej technice widzimy, że grupowanie kolorów win jest mniej wyraźne niż w poprzednim przypadku. Granica między klasami jest trudniejsza do określenia, co prowadzi do częstszych błędów klasyfikacji, szczególnie w obszarze klasy białych win. Widać także, że gdy obserwacje czerwonych win nachodzą głębiej na obszar białych win, model ma tendencję do błędnego klasyfikowania ich jako białe.





W tym przypadku punkty odstające dla obu zmiennych zostały poprawnie sklasyfikowane. Punkty tworzą trzy wyraźne grupy: wina zdecydowanie czerwone, wina mieszane oraz wina zdecydowanie białe. Interesujące jest to, że błędy klasyfikacji pojawiają się wyłącznie w grupach mieszanych i białych win.

W grupie białych win zidentyfikowano jedno czerwone wino o wartości *free.sulfur.dioxide* wynoszącej około 70, które zostało błędnie sklasyfikowane. Większość błędów koncentruje się jednak w grupie mieszanej, co wskazuje na większą trudność modelu w rozróżnieniu obserwacji o pośrednich wartościach cech. To zjawisko podkreśla wyzwanie związane z klasyfikacją win znajdujących się w obszarze granicznym między klasami.



Model popełnia błędy jedynie dla obserwacji o jakości (*quality*) równej 4, 5 i 6. W przypadku jakości równej 4, model dwukrotnie błędnie sklasyfikował białe wino jako czerwone. Dla jakości równej 5 wystąpił po jednym błędzie w obu kierunkach: białe wino zostało sklasyfikowane jako czerwone, a czerwone jako białe.

Najwięcej pomyłek model odnotował dla jakości równej 6. Przy niskiej zawartości alkoholu w tej kategorii model błędnie sklasyfikował białe wino jako czerwone. Kilka błędów wystąpiło także przy poziomie alkoholu około 10.5, a kolejne trzy pomyłki pojawiły się w zakresie zawartości alkoholu od 11 do 12.

Model dobrze radzi sobie z klasyfikacją win, ale napotyka trudności w przypadku obserwacji odstających oraz tych znajdujących się blisko granicy między klasami.

Czerwone wina zazwyczaj wyróżniają się większą gęstością i niższym poziomem cukru resztkowego, co ułatwia ich odróżnienie. Jednak błędy klasyfikacji częściej pojawiają się w obszarach mieszanych cech, gdzie granica między klasami jest mniej wyraźna. Najwięcej pomyłek dotyczy win o średniej jakości (*quality* = 6), szczególnie przy średniej zawartości alkoholu, co podkreśla wyzwanie w rozróżnianiu bardziej zbliżonych obserwacji.

3.2. Model regresji logistycznej na składowych głównych

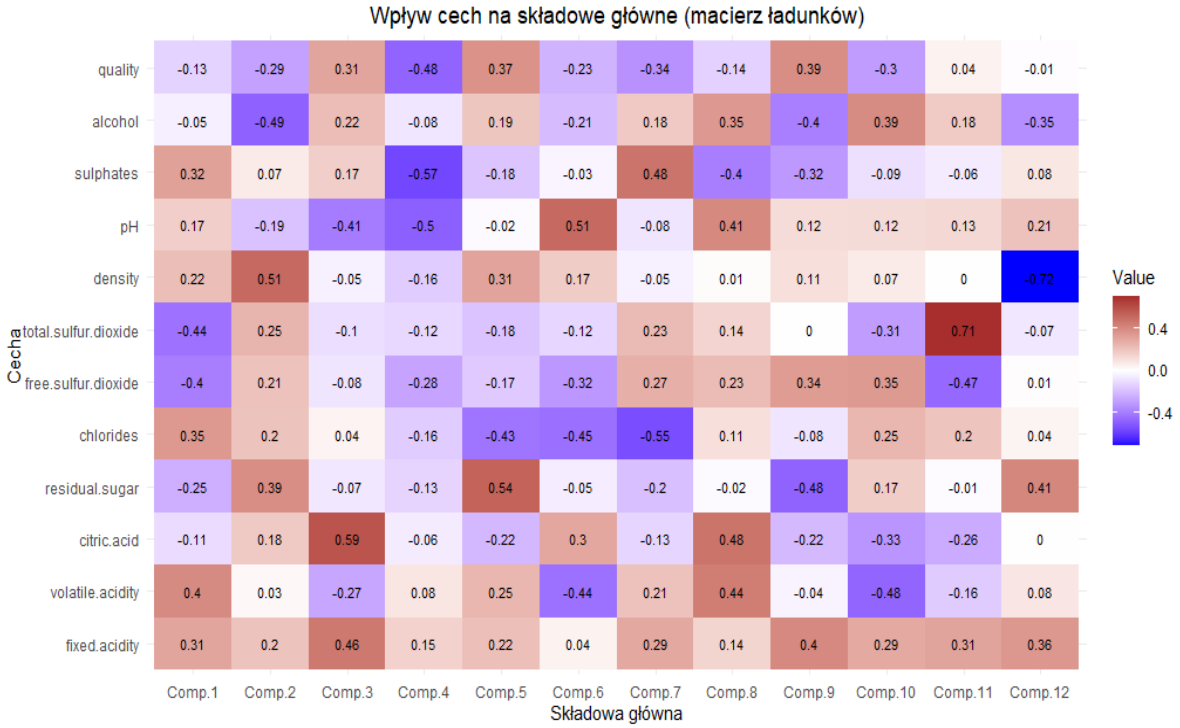
Analiza składowych głównych

Analiza składowych głównych (PCA) to liniowa technika redukcji wymiarowości. Polega na przekształceniu danych do nowego układu współrzędnych, w którym główne składowe odzwierciedlają kierunki o największej zmienności, co umożliwia ich łatwą identyfikację.

PCA przeprowadzamy za pomocą funkcji *princomp* na zbiorze treningowym, z wyłączeniem kolumny docelowej. Dane standaryzujemy, co umożliwia wyrównanie wpływu zmiennych o różnych skalach i zapobiega dominacji tych o większej wariancji. Po przekształceniu dodajemy kolumnę docelową, uzyskując pełny zbiór treningowy.

Dane testowe przekształcamy analogicznie: standaryzujemy je na podstawie średnich i odchyleń standardowych z danych treningowych, a następnie rzutujemy na przestrzeń głównych składowych przy użyciu macierzy ładunków wyznaczonej w PCA zbioru treningowego. Po transformacji dodajemy zmienną docelową, co zapewnia spójność danych testowych z treningowymi i ich gotowość do dalszego modelowania.

Macierz ładunków



Macierz ładunków pokazuje, które cechy mają największy wpływ na poszczególne główne składowe. Intensywność koloru odzwierciedla siłę tego wpływu: wartości dodatnie są przedstawione w odcieniach czerwieni, a wartości ujemne – w odcieniach niebieskiego. Z kolei wartości bliskie 0, oznaczone białym kolorem, wskazują na niewielki wpływ danej cechy na składową.

Dla *Comp.1*, cechy takie jak *volatile.acidity* (0.4) i *chlorides* (0.35) wywierają istotny dodatni wpływ, co oznacza, że dodatnie wartości tych zmiennych są skorelowane z wyższymi wartościami pierwszej składowej głównej. Natomiast zmienne *total.sulfur.dioxide* (-0.44) i *free.sulfur.dioxide* (-0.4) mają istotny ujemny wpływ, sugerując, że wyższe wartości tych zmiennych są skorelowane z niższymi wartościami tej składowej. Podobną analizę można przeprowadzić dla pozostałych składowych głównych, identyfikując, które cechy wywierają największy wpływ na każdą z nich. Pozwala to lepiej zrozumieć, jak zmienne kształtują strukturę danych w przestrzeni głównych składowych.

Wartości w tabeli przedstawiają globalną ważność cech, obliczoną jako sumę wartości absolutnych ładunków dla wszystkich głównych składowych. Im wyższa wartość, tym większy wpływ danej cechy na całościową strukturę danych.

fixed.acidity	***free.sulfur.dioxide***	***alcohol***	***quality***
3.18	3.12	3.09	3.04
volatile.acidity	***Citric.acid***	***pH***	***chlorides***

2.88	2.87	2.85	2.85
sulphates	***residual.sugar***	***total.sulfur.dioxide***	***density***
2.77	2. 73	2.68	2.38

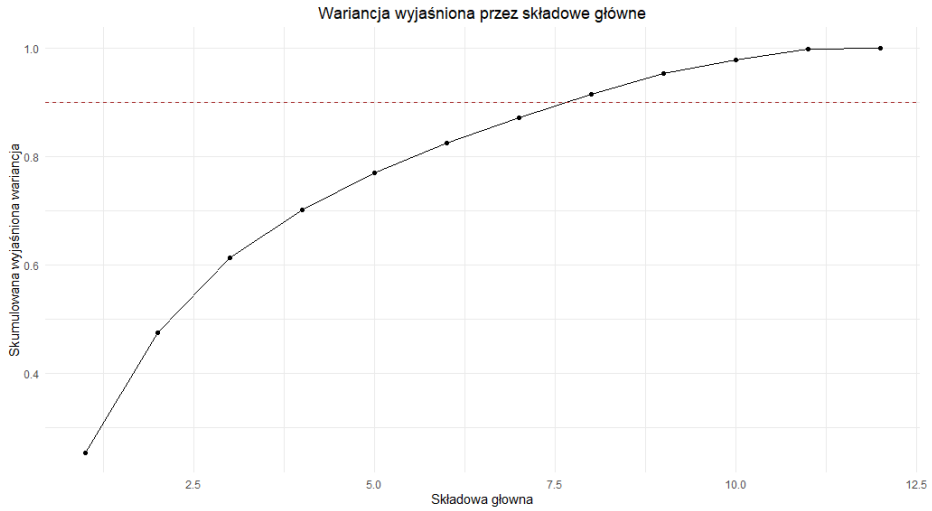
Największy udział w wyjaśnieniu struktury danych mają *fixed.acidity*, *free.sulfur.dioxide, alcohol* oraz quality, co czyni je kluczowymi cechami w analizie. Z kolei *volatile.acidity, citric.acid, pH* oraz *chlorides* wykazują umiarkowany, ale nadal istotny wkład w analizę. *density* ma najniższy wkład, jednak pozostaje cechą, która dostarcza pewnych informacji.

Selekcja składowych głównych

Wyjaśniona wariancja

Wyjaśniona wariancja jest kluczowym elementem *PCA*, który pomaga zrozumieć, jak dobrze dane mogą być opisane w zredukowanej przestrzeni wymiarowej. Wysoka wyjaśniona wariancja oznacza, że dana składowa przechowuje istotne informacje o zmienności danych. Niska wyjaśniona wariancja sugeruje, że składowa zawiera głównie szum lub mniej istotne informacje. *Standard deviation* to miara rozproszenia danych wzdłuż danej składowej głównej.

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6     Comp.7
Standard deviation     1.746041 1.6282769 1.2892563 1.03001930 0.90101398 0.81697380 0.74024236
Proportion of Variance 0.254055 0.2209405 0.1385152 0.08841165 0.06765218 0.05562052 0.04566323
Cumulative Proportion  0.254055 0.4749955 0.6135106 0.70192227 0.76957446 0.82519497 0.87085820
                           Comp.8     Comp.9    Comp.10    Comp.11     Comp.12
Standard deviation     0.72107734 0.68232765 0.55641613 0.47203147 0.178228486
Proportion of Variance 0.04332938 0.03879759 0.02579991 0.01856781 0.002647116
Cumulative Proportion  0.91418758 0.95298517 0.97878508 0.99735288 1.000000000
```

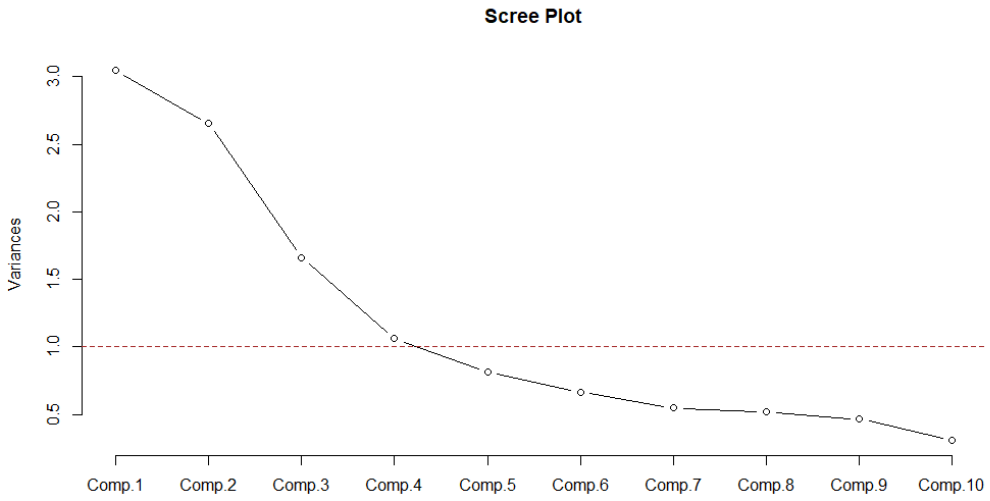


Na powyższym wykresie możemy łatwo zauważyć, że pierwszych osiem składowych głównych wyjaśnia ponad 90% wariancji. Oznacza to, że większość informacji zawartych w oryginalnym zbiorze danych może zostać zachowana przy użyciu

jedynie tych ośmiu składowych, co znacznie redukuje wymiarowość danych przy minimalnej utracie informacji.

Wykres osypiska i kryterium *Kaisera*

Wykres osypiska (*scree plot*) przedstawia wartości własne poszczególnych składowych w malejącej kolejności. Na wykresie można zaobserwować "kolano", czyli punkt, w którym wartości własne zaczynają maleć w wolniejszym tempie. Składowe znajdujące się przed tym punktem są uznawane za istotne, ponieważ wyjaśniają znaczną część zmienności w danych. Kryterium *Kaisera* z kolei sugeruje zachowanie wyłącznie tych składowych, których wartości własne są większe niż 1. Oznacza to, że każda z tych składowych wyjaśnia więcej wariancji niż pojedyncza oryginalna zmienna.



Powyżej przedstawiono wykres osypiska dla analizowanych danych, na którym czerwona linia oznacza wartość własną równą 1, zgodnie z kryterium Kaisera. Na podstawie wykresu można zauważyć, że cztery pierwsze składowe główne spełniają to kryterium, co sugeruje, że są one wystarczające, aby dobrze odzwierciedlić pierwotne dane, zachowując większość zawartej w nich zmienności.

AIC (Akaike Information Criterion)

StepAIC to funkcja dostępna w pakiecie *MASS*, służąca do automatycznego wyboru najlepszego modelu statystycznego przy użyciu kryterium *AIC*. Funkcja przeprowadza proces selekcji zmiennych w modelu, wykorzystując regresję krokową.

```
Call:
glm(formula = type ~ Comp.1 + Comp.3 + Comp.4 + Comp.5 + Comp.8 +
    Comp.9 + Comp.10 + Comp.11 + Comp.12, family = "binomial",
    data = train_scores)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.5034     0.2630 -13.319  < 2e-16 ***
Comp.1        3.6262     0.2291  15.826  < 2e-16 ***
Comp.3        0.3207     0.1630   1.968 0.049093 *
Comp.4       -1.0274     0.1813  -5.665 1.47e-08 ***
Comp.5        1.0450     0.2426   4.307 1.65e-05 ***
Comp.8        0.8690     0.2619   3.318 0.000908 ***
Comp.9        1.1385     0.2866   3.973 7.10e-05 ***
Comp.10       1.3554     0.3530   3.840 0.000123 ***
Comp.11      -2.2040     0.4311  -5.113 3.17e-07 ***
Comp.12      -7.7786     1.0620  -7.324 2.40e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4234.38  on 3724  degrees of freedom
Residual deviance:  237.84  on 3715  degrees of freedom
AIC: 257.84

Number of Fisher Scoring iterations: 9
```

Po zastosowaniu funkcji dla naszych danych, otrzymujemy model, który uwzględnia następujące główne składowe: Comp.1, Comp.3, Comp.4, Comp.5, Comp.8, Comp.9, Comp.10, Comp.11 oraz Comp.12.

Aby wybrać odpowiednie komponenty do końcowego modelu regresji logistycznej opartego na PCA, przeprowadzamy ocenę predykcji modeli wybranych według powyższych kryteriów na zbiorze treningowym. Proces ten polega na podziale danych treningowych na trzy części, z których każda kolejno pełni rolę zbioru testowego dla jednego z trzech modeli, podczas gdy pozostałe dwie części służą jako zbiór treningowy. Taka procedura pozwala ocenić, jak modele radzą sobie z nowymi, nieznanymi danymi, jednocześnie zachowując oryginalny zbiór testowy w nienaruszonym stanie do ostatecznej weryfikacji wyników.
Otrzymujemy następujące wyniki:

Wyjaśniona wariancja	**Kryterium Kaisera**	**AIC**
0.9887188	0.983884	0.9943685

Analiza dokładności modeli na zbiorze treningowym wykazała, że model wybrany na podstawie kryterium *AIC* osiągnął najwyższą dokładność, z różnicą 0.0056 względem modelu opartego na wyjaśnionej wariancji i około 0.01 względem modelu zgodnego z kryterium *Kaisera*. Model *AIC* wykorzystuje 9 komponentów, model wyjaśnionej wariancji 8, a model *Kaisera* tylko 4. Choć różnice w dokładności są niewielkie, prostsze modele z mniejszą liczbą komponentów są bardziej intuicyjne i mniej podatne na przeuczenie, co przemawia za wyborem modelu o najmniejszej złożoności.

Opis modelu

Przy finalnym modelu regresji logistycznej na składowych głównych używamy komponentów wybranych przez kryterium *Kaisera*.

```
Call:
glm(formula = type ~ ., family = "binomial", data = train_scores[,
    c(components_kaiser_names, "type")])

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.21262    0.24842 -16.958  < 2e-16 ***
Comp.1       3.85880    0.20499  18.825  < 2e-16 ***
Comp.2      -0.25652    0.09951  -2.578  0.00994 **
Comp.3       0.19585    0.09721   2.015  0.04393 *
Comp.4      -0.91045    0.12162  -7.486 7.12e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4234.38  on 3724  degrees of freedom
Residual deviance:  454.95  on 3720  degrees of freedom
AIC: 464.95

Number of Fisher Scoring iterations: 9
```

Wnioski z modelu

Znak współczynników:

Comp.1 ma największy dodatni wpływ (wzrost jego wartości znacząco zwiększa szansę na przewidywanie klasy 1 – czerwonego wina). *Comp.2* i *Comp.4* mają ujemny wpływ (wzrost ich wartości zmniejsza szanse przypisania do klasy 1). *Comp.3* ma niewielki dodatni wpływ, ale słabszy w porównaniu do reszty predyktorów.

Istotność zmiennych:

Wszystkie wybrane komponenty są istotne statystycznie ($p < 0.05$). Najbardziej istotnym zdaje się być *Comp.1*.

***Null deviance* i *residual deviance*:**

4234.38 - poziom zmienności w danych dla modelu z samym interceptem.

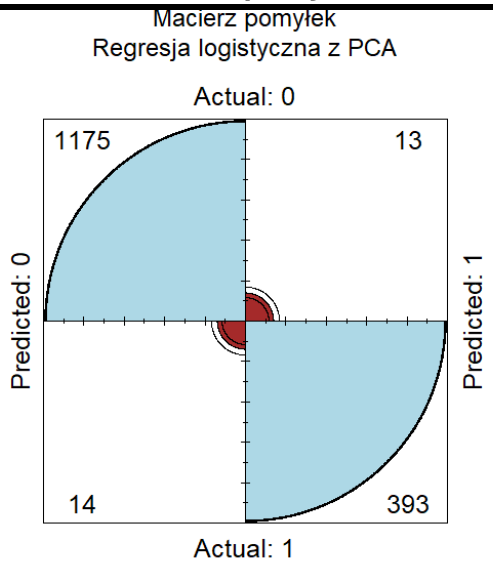
454,95 - zmienność w danych po uwzględnieniu wybranych komponentów jako predyktorów. Znacząca redukcja w *deviance* wskazuje, że model dobrze dopasowuje się do danych.

Iteracja Fisher Scoring:

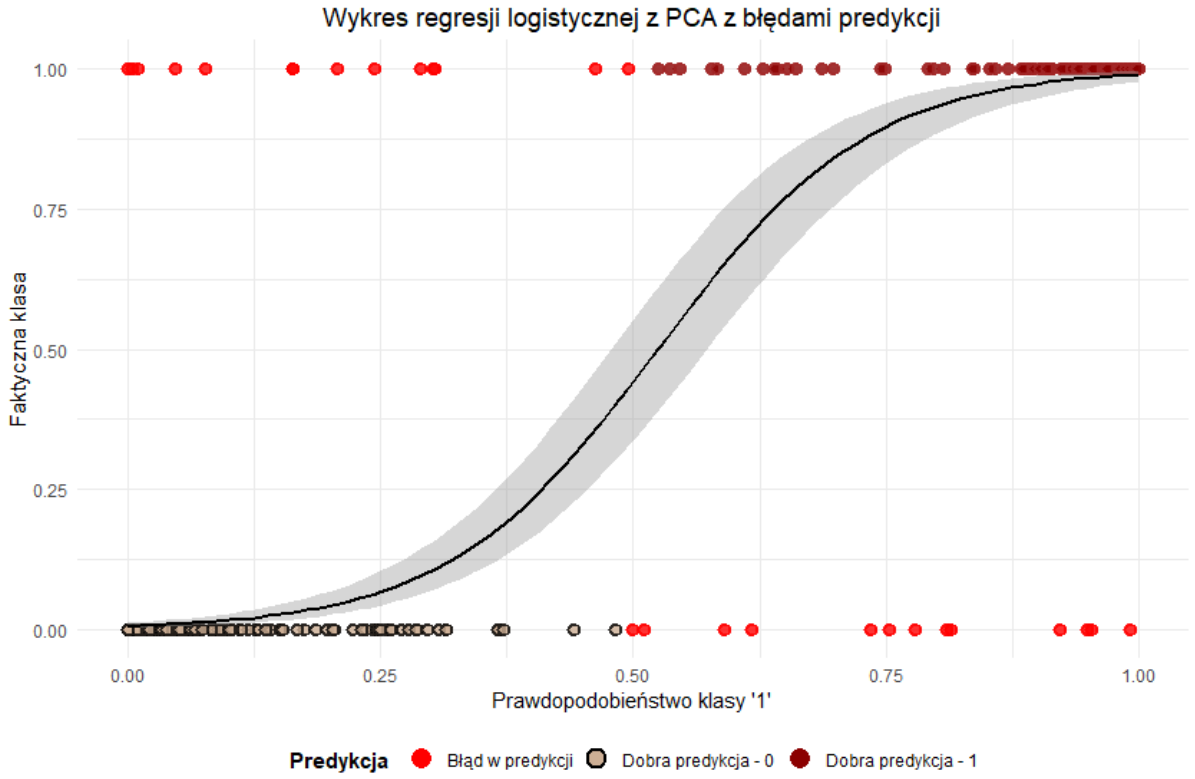
Model zbiega się po 9 iteracjach algorytmu *Fishera*, co wskazuje na stabilność i poprawność dopasowania.

Ocena jakości modelu

Macierz pomyłek

Macierz pomyłek	Dokładność modelu
	98.30721%

Wykres krzywej sigmoidalnej

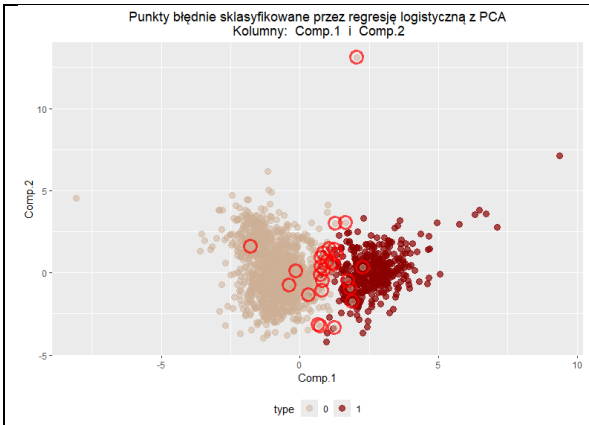


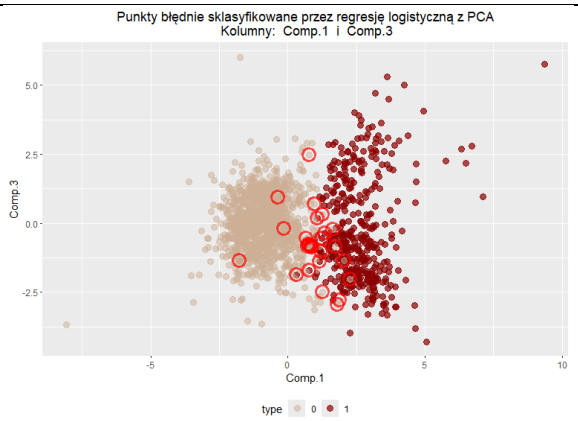
Czy model jest dobry?

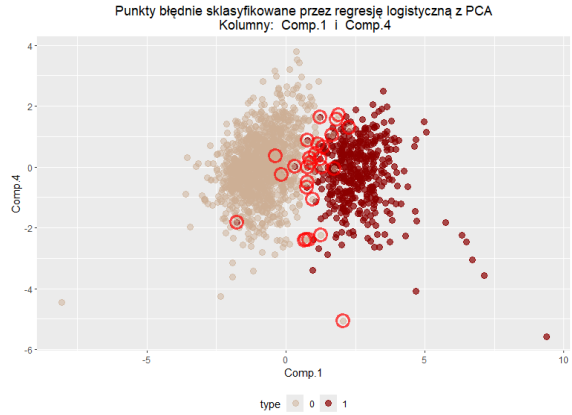
1. **Wysoka dokładność na zbiorze testowym** - model poprawnie klasyfikuje 98,4% obserwacji, co świadczy o wysokiej skuteczności ogólnej.
2. **Niewielka liczba błędów** - model myli się w zaledwie 27 przypadkach na 1595, osiągając imponujące wyniki w klasyfikacji obu klas.
3. **Istotność zmiennych** - wszystkie wybrane komponenty są istotne statystycznie ($p < 0.05$), co potwierdza ich znaczenie dla wyjaśniania zmienności w danych.
4. **Dopasowanie a złożoność** - model opiera się na 4 komponentach, zapewniając równowagę między prostotą a skutecznością.
5. ***Deviance*** - znaczący spadek *Null Deviance* do *Residual Deviance* pokazuje, że model skutecznie redukuje niedopasowanie, dobrze wyjaśniając zmienność w danych.

Analiza błędów

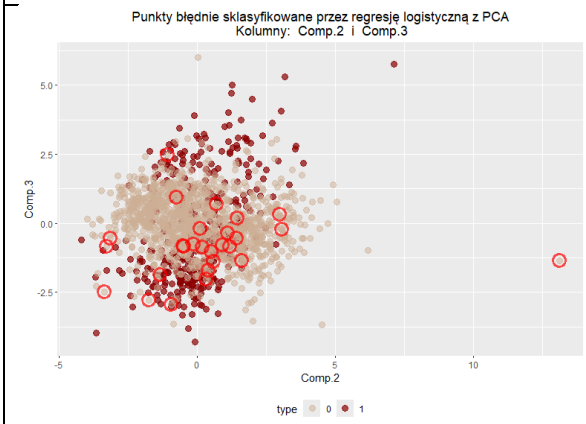
Analizę błędów przeprowadzimy w sposób analogiczny jak dla klasycznej regresji logistycznej, z tą różnicą, że w modelu z *PCA* zmiennymi są wybrane główne komponenty. Jasne punkty na wykresie reprezentują wina białe, natomiast bordowe punkty wina czerwone. Błędne klasyfikacje zostały oznaczone czerwonymi okręgami dla wygodnej wizualizacji.

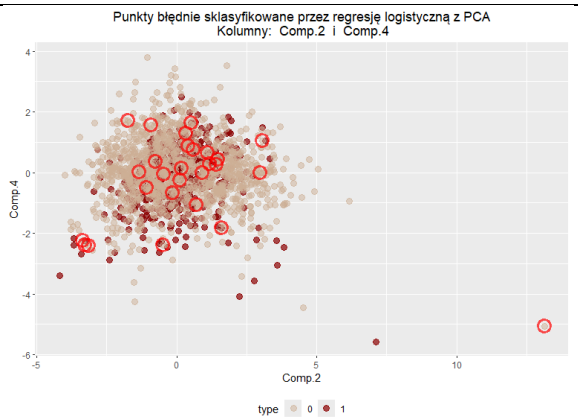




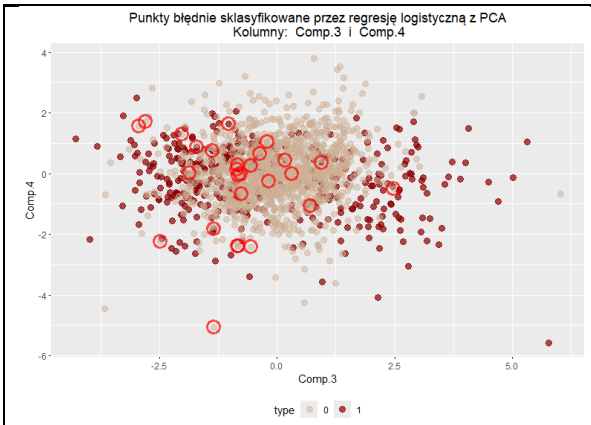


Na wykresach przedstawiających błędne klasyfikacje w przestrzeni *Comp.1* i pozostałych komponentów można zauważyć, że klasy wina są w dużej mierze rozseparowane. Może to sugerować, że cechy chemiczne i sensoryczne win użyte do obliczeń PCA są istotnie związane z różnicami między klasami, a *Comp.1,* jako komponent wyjaśniający największą część wariancji, jest kluczowy dla separacji tych klas. Należy jednak pamiętać, że PCA nie bierze pod uwagę klasy docelowej podczas tworzenia komponentów, a separacja wynika jedynie z naturalnych różnic w danych. Można również zaobserwować, że model popełnia błędy w przypadkach, gdzie wartości dla obu klas są zbliżone, co jest typowe dla tego typu modeli.





Dane przedstawione na wykresach pokazują błędne klasyfikacje w przestrzeni wyznaczonej przez *Comp.2, Comp.3 i Comp.4*. Można zauważyć, że separacja klas jest mniej wyraźna niż w przypadku *Comp.1*. Wynika to z charakterystyki PCA, gdzie każdy kolejny komponent wyjaśnia coraz mniejszą część całkowitej wariancji danych. W przestrzeni komponentów *Comp.2, Comp.3 i Comp.4* różnice między klasami są mniej widoczne, co prowadzi do większego nakładania się punktów należących do różnych klas. Błędne klasyfikacje koncentrują się w obszarach, gdzie klasy nakładają się na siebie, co wskazuje na trudności modelu w rozróżnianiu obserwacji o podobnych wartościach komponentów. Dodatkowo można zauważyć, że model myli się w przypadku wartości odstającej dla *Comp.2* (prawa dolna strefa wykresów).



Wykres przedstawia błędne klasyfikacje w przestrzeni wyznaczonej przez *Comp.3 i Comp.4,* gdzie separacja klas wina jest jeszcze słabsza niż w przypadku innych komponentów. *Comp.3* i *Comp.4* wyjaśniają zmienność, która nie jest bezpośrednio związana z różnicami między klasami, co ogranicza ich przydatność w procesie klasyfikacji. Model napotyka trudności w skutecznym wykorzystaniu tych komponentów, ponieważ dostarczają one niewystarczających informacji różnicujących klasy. Błędne klasyfikacje koncentrują się w obszarach, gdzie wartości obu komponentów nakładają się na siebie.

3.3. Model Relaxed LASSO

Opis modelu

Zastosowaliśmy metodę Relaxed LASSO do budowy modelu regresji logistycznej. Ta technika pozwala na jednoczesne przeprowadzenie selekcji zmiennych oraz regularyzcję modelu, co ułatwia uproszczenie modelu i poprawia jego dopasowanie do danych. Model został dopasowany przy użyciu 10-krotnej walidacji krzyżowej, co pozwala na ocenę jego stabilności i skuteczności na różnych podzbiorach danych.

Wyniki kroswalidacji:

Measure: Binomial Deviance							
	Gamma	***Index***	***Lambda***	***Index***	***Measure***	***SE***	***Nonzero***
min	0	1	0.000493	70	0.06652	0.01565	12
1se	0	1	0.026923	27	0.08155	0.01198	8

Parametr *Gamma* = 0 oznacza, że model wykorzystuje współczynniki wyznaczone w pierwszym etapie regresji LASSO, bez wprowadzania dodatkowego "rozluźnienia".

Model z minimalnym błędem *Binomial Deviance* (*lambda* = 0.000493) jest bardziej złożony i zawiera 12 zmiennych z niezerowymi współczynnikami. To sugeruje, że model lepiej dopasowuje się do danych, ale może być bardziej podatny na przeuczenie.

Z kolei model uwzględniający błąd standardowy *1-se* (*lambda* = 0.026923) jest prostszy, zawiera tylko 8 zmiennych z niezerowymi współczynnikami. Jego dopasowanie jest nieco gorsze, a błąd *Binomial Deviance* wyższy, ale jest to bardziej konserwatywny model, który może być odporniejszy na przeuczenie.

Biorąc pod uwagę, że w poprzednich przypadkach regresji logistycznej oraz regresji logistycznej na *PCA* zwracaliśmy szczególną uwagę na złożoność modeli oraz ich podatność na przeuczenie, zdecydowaliśmy się wybrać model z parametrem *1-se*, który jest prostszy i bardziej odporny na przeuczenie.

Współczynniki

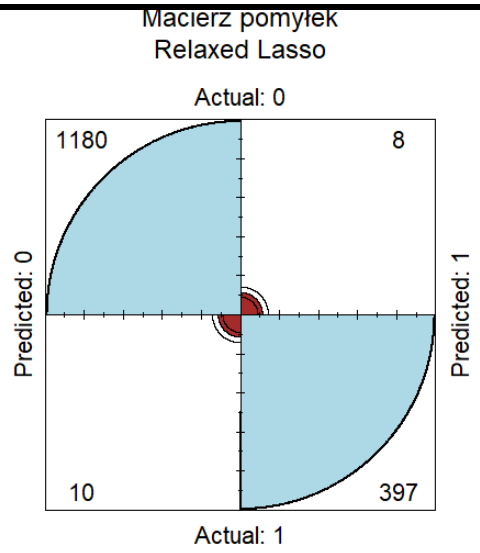
fixed.acidity	***volatile.acidity***	***residual.sugar***	***chlorides***
0.32338022	8.76201038	-0.31857394	18.53112464
total.sulfur.dioxide	***density***	***pH***	***sulphates***
-0.05709239	907.24561377	5.31842654	8.76152571
citric.acid	***free.sulfur.dioxide***	***alcohol***	***quality***
0	0	0	0

Aż 6 z 8 zmiennych posiada dodatni współczynnik, co wskazuje, że ich wzrost zwiększa logarytmiczny iloraz szans (*log-odds*) przypisania próbki do klasy 1 – wina czerwonego. Z kolei dwie zmienne mają ujemny współczynnik, co oznacza, że ich wzrost zmniejsza *log-odds* przypisania próbki do tej klasy.

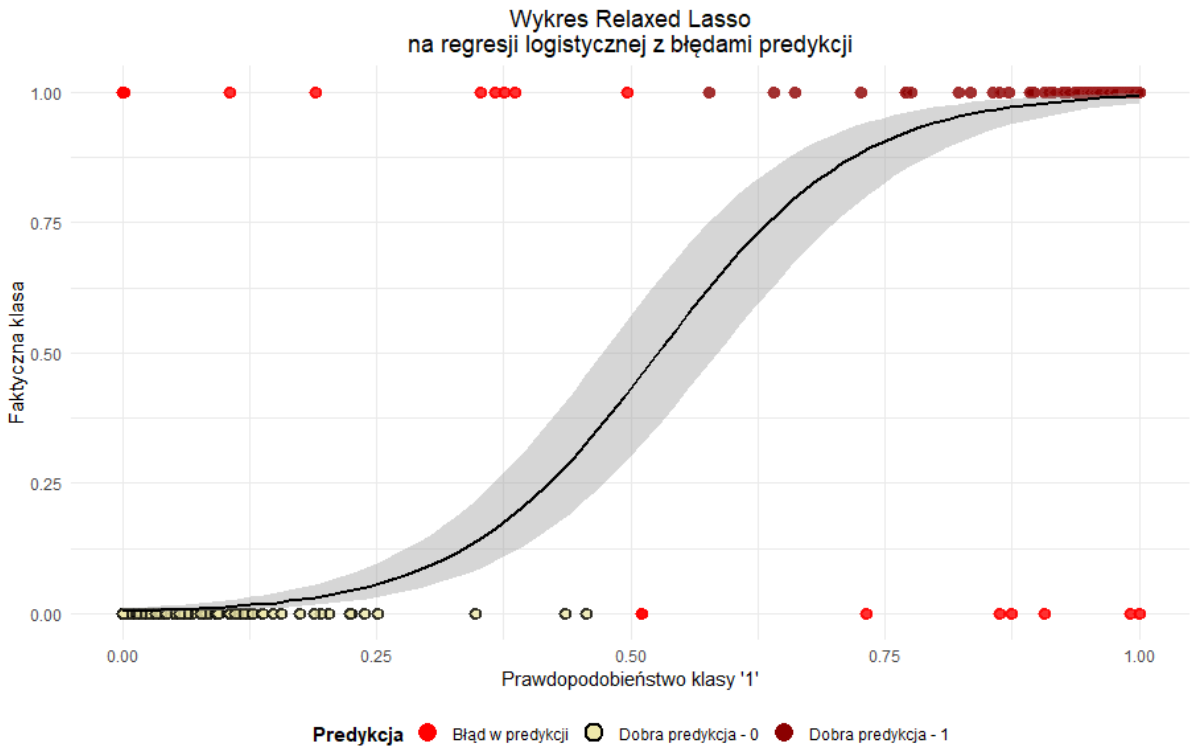
Cztery zmienne, które zostały wykluczone z modelu, to: kwas cytrynowy, wolny dwutlenek siarki, zawartość alkoholu oraz ocena jakości.

Ocena jakości modelu

Macierz pomyłek

Macierz pomyłek	**Dokładność modelu**
	98.87147%

Wykres krzywej sigmoidalnej

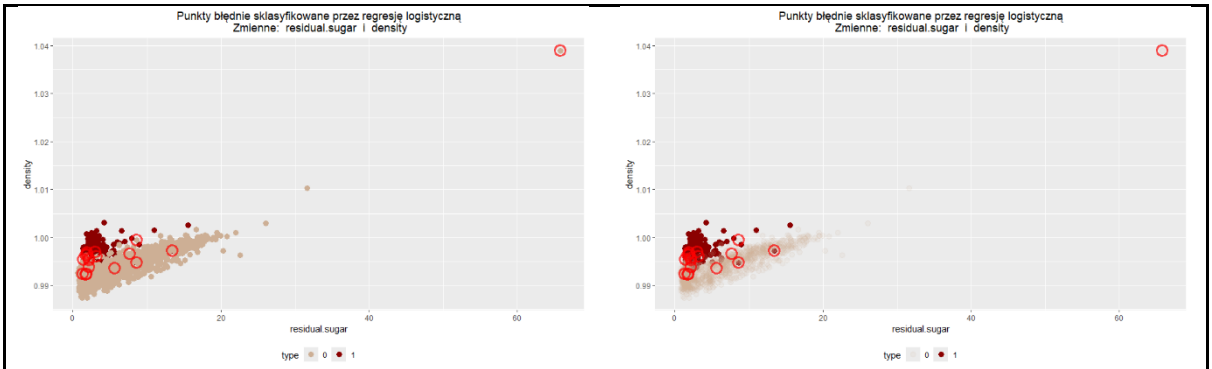


Czy model jest dobry?

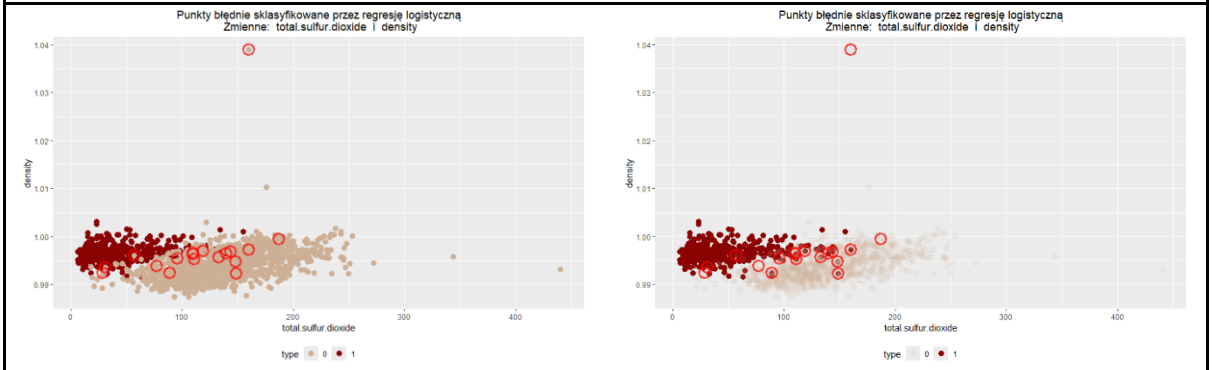
Model można uznać za dobry. Dzięki zastosowaniu metody Relaxed LASSO udało się jednocześnie zredukować liczbę zmiennych i poprawić dopasowanie modelu, co sprzyja jego interpretowalności i stabilności. Wybrany prostszy wariant modelu (*1-se*) osiąga dobrą równowagę między dokładnością a odpornością na przeuczenie. Wysoka dokładność klasyfikacji na poziomie 98,87% dodatkowo potwierdza skuteczność modelu w rozróżnianiu klas. Wykluczenie mniej istotnych zmiennych, takich jak zawartość alkoholu czy ocena jakości, wskazuje na to, że model jest oparty na kluczowych cechach istotnych dla klasyfikacji, co wzmacnia jego wartość predykcyjną.

Analiza błędów

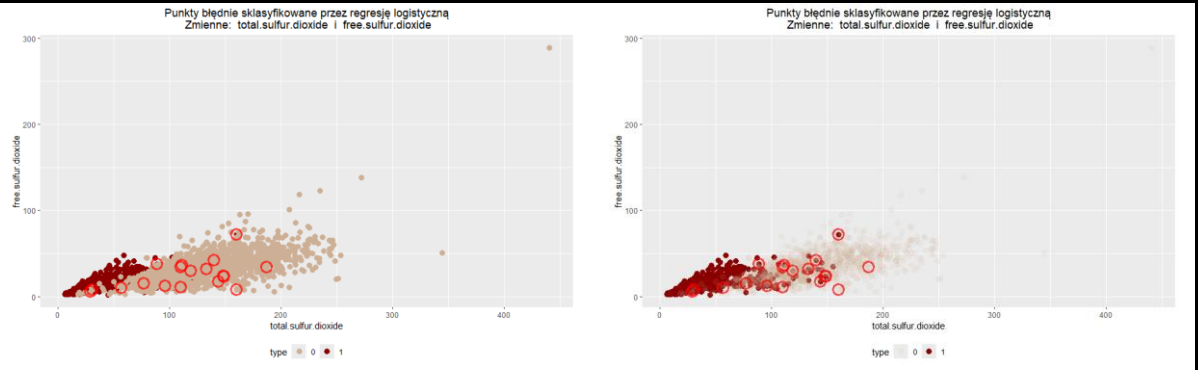
Podobnie jak w poprzednich modelach, analizę błędów przedstawimy za pomocą wykresów punktowych, na których zaznaczone są błędne klasyfikacje. Sposób oznaczania punktów i błędów jest identyczny jak w poprzednich analizach.



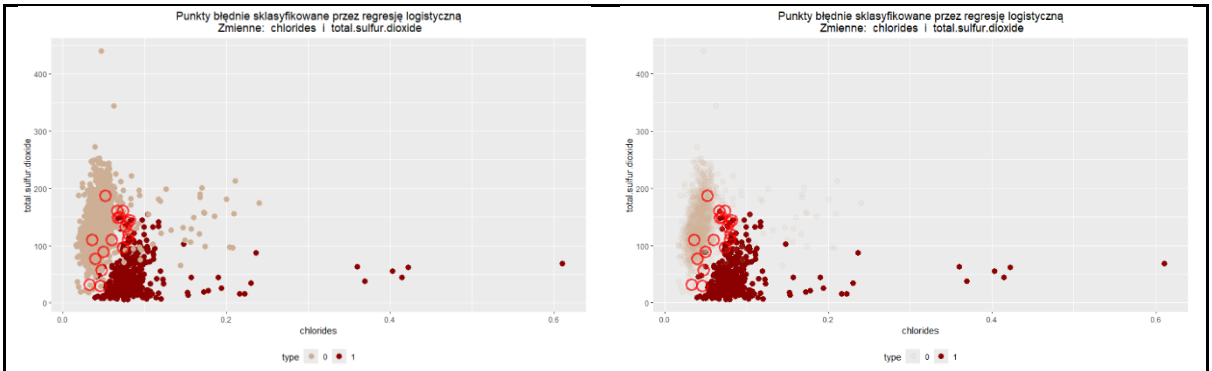
Model popełnił błąd dla największego *outliera*, jednak pozostałe obserwacje odstające zostały poprawnie sklasyfikowane. Wiele błędów występuje po skrajnie lewej stronie gęsto skonsolidowanej grupy punktów, gdzie próbki dla zmiennej *residual.sugar* mają wartości w zakresie od 0 do 5. Dodatkowo, model popełnił kilka błędów, ponieważ czerwone wino weszło w obszar zarezerwowany dla białych win i odwrotnie.



Model ponownie popełnił błąd w przypadku największego *outliera* zmiennej *density*. Natomiast dla wartości odstających drugiej zmiennej model poradził sobie bezbłędnie. Problemy pojawiają się w obszarze, gdzie grupy czerwonych i białych win nakładają się na siebie. Wiele błędów klasyfikacji występuje w strefie, w której oba rodzaje wina mieszają się, co utrudnia modelowi poprawne przypisanie klasy.



Model radzi sobie bezbłędnie z wartościami odstającymi obu zmiennych. Jednak błędy pojawiają się w obszarze, gdzie oba rodzaje win się mieszają. W szczególności czerwone wina znajdujące się głęboko w obszarze białego wina stanowiły dla modelu coraz większe wyzwanie, prowadząc do częstszych pomyłek.



W tym przypadku model popełnia błędy, gdy zmienna *chlorides* przyjmuje wartości z zakresu 0.033–0.082. Przy wyższych wartościach model radzi sobie z klasyfikacją poprawnie, nawet w przypadku obserwacji odstających. Zdecydowana większość błędów w podanym zakresie występuje w obszarze, gdzie dochodzi do mieszania się win białych i czerwonych. Potwierdza to, że model ponownie napotyka trudności w rozróżnianiu obserwacji w takim środowisku.

Model ma trudności z poprawną klasyfikacją w obszarach, gdzie klasy nakładają się na siebie lub gdzie występują nietypowe wzorce danych. Wartości odstające zazwyczaj nie sprawiają mu problemu, ale gdy próbki różnych klas znajdują się w bliskim sąsiedztwie, model staje się mniej precyzyjny. Większość błędów występuje w strefach przejściowych między grupami danych, co wskazuje na wyzwania związane z rozróżnianiem obserwacji w takich obszarach.

4. Porównanie modeli

Zestawienie zalet i wad analizowanych modeli

MODEL	ZALETY	WADY
Regresja logistyczna	• Uproszczenie w stosunku do pełnego modelu poprzez selekcję zmiennych; • Wysoka dokładność na zbiorze testowym;	• Ryzyko przeuczenia z powodu braku regularyzacji i dużej liczby wykorzystywanych zmiennych objaśniających; • Wrażliwość na dane odstające;
Regresja logistyczna z wykorzystaniem analizy składowych głównych	• Redukcja wymiarowości poprzez analizę składowych głównych (PCA); • Zwiększona odporność na dane zawierające zależności między	• Trudniejsza interpretacja wyników, ponieważ składowe nie odnoszą się bezpośrednio do pierwotnych zmiennych;

	zmiennymi objaśniającymi; • Wyższa odporność na przeuczenie w porównaniu ze zwykłym podejściem; • Wysoka dokładność na zbiorze testowym;	• Możliwe straty informacji w wyniku redukcji wymiarowości; • Zwiększona złożoność obliczeniowa;
Regresja logistyczna z wykorzystaniem metody relaxed LASSO	• Wprowadzenie reguralyzacji (LASSO), która zwiększa odporność na przeuczenie; • Uproszczenie modelu poprzez selekcję zmiennych; • Wysoka dokładność na zbiorze testowym;	• Wrażliwość na wybór parametru lambda – zbyt duża regularyzacja może nadmiernie uprościć model; • Zależność od optymalizacji parametru regularyzacji lambda;

Wymogi i złożoność a dokładność

MODEL	WYMOGI	LICZBA PREDYKTORÓW	DOKŁADNOŚĆ
Regresja logistyczna	• Wymaga procesu selekcji zmiennych objaśniających;	**10**	**99.18%**
Regresja logistyczna z wykorzystaniem analizy składowych głównych	• Wymaga redukcji wymiarowości, co dodaje krok do przetwarzania danych; • Wymaga selekcji składowych do użycia w modelu;	**4** (składowe główne)	**98.31%**
Regresja logistyczna z wykorzystaniem metody relaxed LASSO	• Wymaga optymalizacji parametru reguralizacji lambda;	**8**	**98.87%**

Który model jest najlepszy?

Wybór najlepszego modelu zależy od celów analizy i priorytetów.

- Najwyższą dokładność osiąga klasyczna regresja logistyczna (99.18%), co czyni ją najlepszym wyborem w przypadku, gdy głównym celem jest maksymalizacja skuteczności predykcji. Należy jednak uwzględnić potencjalne ryzyko przeuczenia związane z brakiem regularyzacji.
- Relaxed LASSO (*1-se*) oferuje optymalne połączenie prostoty i interpretowalności, wybierając kluczowe zmienne (8) oraz zachowując wysoką dokładność predykcji (98.87%). Jest to model szczególnie rekomendowany w przypadku, gdy wymagana jest czytelność wyników oraz odporność na przeuczenie.
- Regresja logistyczna z *PCA* sprawdzi się najlepiej, gdy celem jest redukcja wymiarowości danych, zwłaszcza w przypadku dużych zbiorów danych. Osiągnięta dokładność (98.31%) pozostaje wysoka, choć trudności interpretacyjne wynikające z użycia analizy głównych składowych mogą ograniczać jej zastosowanie w bardziej praktycznych scenariuszach.

Dla naszych danych najlepszym wyborem wydaje się być **klasyczna regresja logistyczna**, która oferuje najwyższą dokładność przy umiarkowanej złożoności przygotowania modelu. Selekcja zmiennych przeprowadzana w ramach tego modelu skutecznie upraszcza analizę, przy jednoczesnym zachowaniu doskonałej skuteczności predykcyjnej.