Statystyczna Analiza Danych – Projekt 2

Machine learning a prognozowanie

Paweł Brodziak

1. Wstęp

Celem poniższej pracy jest wykonanie predykcji z użyciem metod nauczania maszynowego, analiza jakości wykonanych klasyfikacji oraz porównanie użytych metod. W pracy wykonane będą również porównania metod i parametrów w nich użytych oraz ocena który zestaw parametrów jest najbardziej odpowiedni. Badanym zbiorem danych będą dane 'loan' które opisują parametry kredytobiorców oraz obecny status ich kredytu (tak/nie) który będzie zmienną opisywaną.

2. Opis danych

Wybrany zbiór danych zawiera następujące informacje o kredytobiorcach:

- Gender - płeć: Mężczyzna (0) lub kobieta (1)
- Married - małżeństwo: Tak(1) lub Nie (0)
- Dependents – liczba osób na utrzymaniu
- Education – edukacja: wyższe wykształcenie (1) lub jego brak (0)
- Self_Employed – samozatrudnienie: tak (1) lub nie (0)
- ApplicantIncome – dochód wnioskującego
- CoapplicantIncomey – dochód współwnioskującgeo
- LoanAmount – kwota kredytu
- Loan_Amount_Term – czas trwania kredytu
- Credit_History – historia kredytowa: pozytywna(1) lub negatywna (0)
- Property_Area – obszar znajdowania się nieruchomości Urban/Semiurban/Rural
- Loan_Status – status kredytu Y/N

Podstawowe statystyki na temat danych

	Gender	Marri	Dependents	Educatior	Self_Employe	ApplicantInc	CoapplicantI	LoanAmou	Loan_Amount	Credit_His	Property	Loan_S
count	614.00	614.00	614.00	614.00	614.00	614.00	614.00	614.00	614.00	614.00	614.00	614.00
mean	0.19	0.65	0.76	0.78	0.14	5403.46	1621.24	146.91	342.12	0.84	1.04	0.69
std	0.39	0.48	1.01	0.41	0.35	6109.04	2926.25	85.55	64.76	0.36	0.79	0.46
min	0.00	0.00	0.00	0.00	0.00	150.00	0.00	9.00	12.00	0.00	0.00	0.00
25%	0.00	0.00	0.00	1.00	0.00	2877.50	0.00	100.00	360.00	1.00	0.00	0.00
50%	0.00	1.00	0.00	1.00	0.00	3812.50	1188.50	128.00	360.00	1.00	1.00	1.00
75%	0.00	1.00	2.00	1.00	0.00	5795.00	2297.25	166.75	360.00	1.00	2.00	1.00
max	1.00	1.00	3.00	1.00	1.00	81000.00	41667.00	700.00	480.00	1.00	2.00	1.00

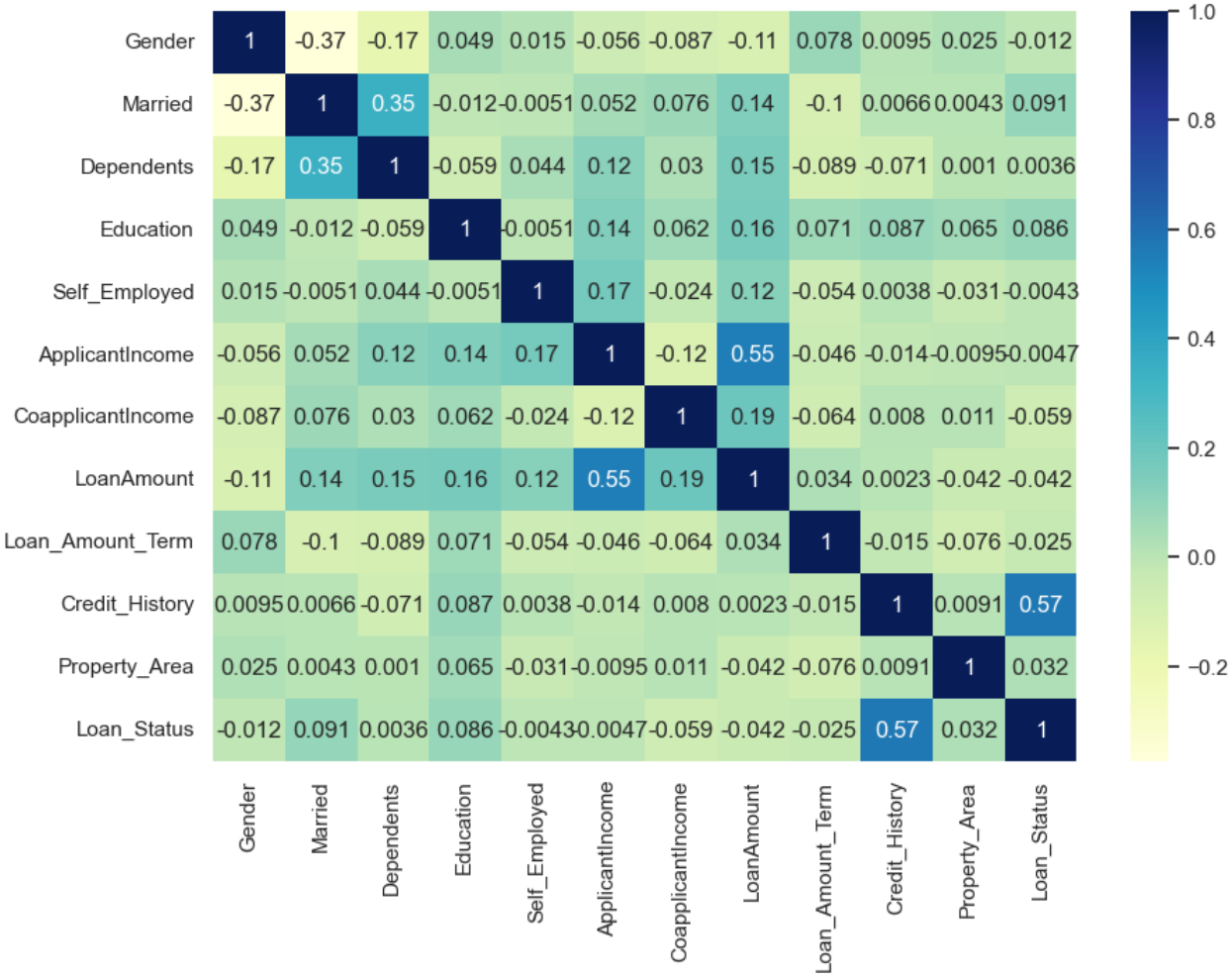
Patrząc na powyższcze statystyki można zauważyć, że większość kredytów została udzielona. Około 69% wniosków ma status pozytywny, chociaż jeszcze większy procent wnioskujących ma pozytywną historię kredytową - 84%.

Można również wywnioskować, że tylko 19% wnioskujących to kobiety,a więc o kredyty znacznie częściej wnioskowali mężczyźni. Ponad połowa wnioskujących była żonata/zamężna (65%). Średnio mieli 0,76 osób na utrzymaniu chociaż ponad połowa osób nie miała żadnej takiej osoby. Maksymalna liczba osób na utrzymaniu wynosiła 3.

78% wnioskujących ma wyższe wykształcenie, a jedynie 14% prowadziło własną działalność gospodarczą.

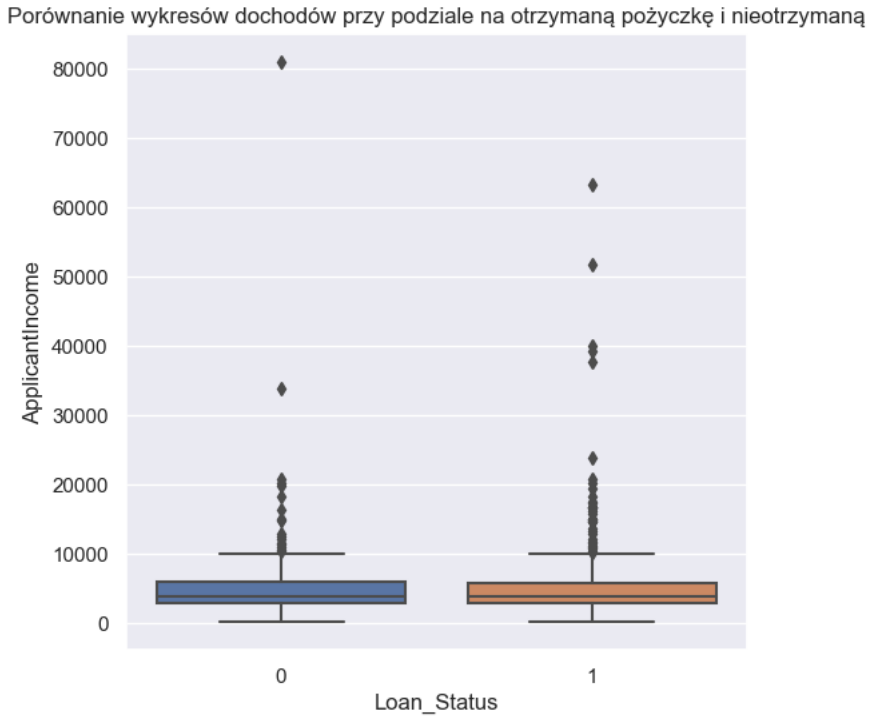
Średni dochód aplikanta wyniósł 5403 a średni dochód współwnioskodawcy 1621. Jednakże średnia może być zaniżona ponieważ w ponad 25% przypadków dochód ten wynosił 0 (z powodu braku współwnioskodawcy lub braku dochodu). Obie kwoty mają bardzo wysokie odchylenie wynoszące kolejno 6109 i 2926. Wynika ono z wysokich wartości maksymalnych. Można stwierdzić, że dla kwót dochodów istnieją duże wartości odstające.

Średnią kwotą wnioskowanego kredytu było 146.90 przy średnim czasie trwania kredytu 342 miesięcy. Jednakże ponad 75% wnioskujących wybierała kredy na 360 miesięcy lub dłużej



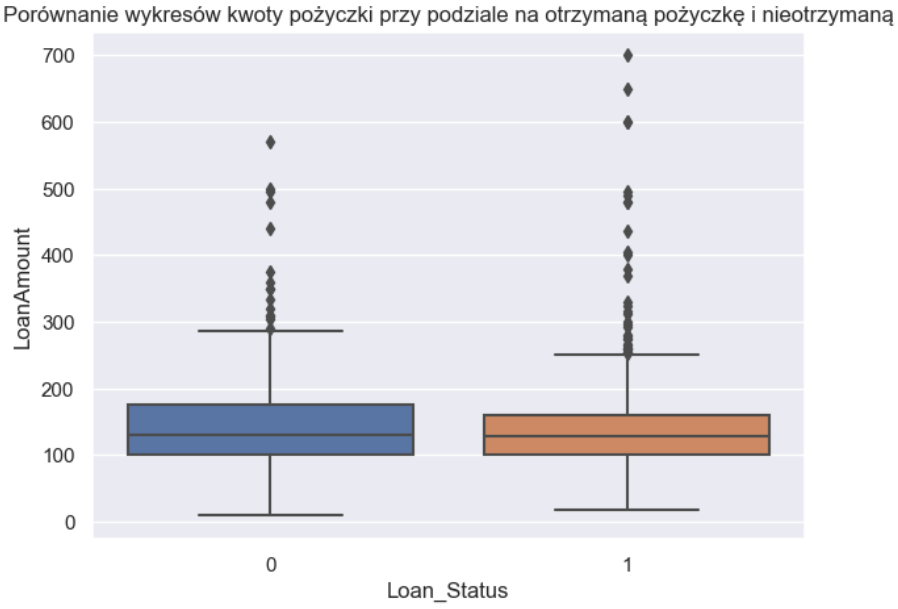
Jak możemy zauważyć na macierzy korelacji, prawie wszystkie zmienne wykazują niemal zerową korelację z naszą zmienną objaśnianą – statusem kredytu. Jedynym wyjątkiem jest tutaj historia kredytowa która jest z nią skorelowana dodatnio w średnim stopniu.

Większość zmiennych nie wykazuje również korelacji między sobą. Jedynie wartości dochodu oraz kwoty wnioskowanego kredytu są ze sobą średnio dodatnie skorelowane oraz zmienne określające małżeństwo i ilość osób na utrzymaniu. Korelacje te są zrozumiałe. Im wyższe dochody tym większe możliwości i potrzeby kredytowe. W przypadku osób na utrzymaniu, zwykle występują one w małżeństwach.



Korzystając z powyższego wykresu pudełkowego możemy potwierdzić nasze wcześniejsze przypuszczenie. Dla zmiennej dochodu wnioskodawcy występują silne dodatnie wartości odstające. Jednak nie są one skorelowane ze statusem kredytu. Średnie oraz odchylenia dochodów w grupie która otrzymała kredyt są zbliżone do wartości grupy która go nie otrzymała.

Dla zmiennej kwoty kredytu, której wykres znajduje się poniżej możemy wyciągnąć takie same wnioski. Średnia, mediana i odchylenia kwot kredytu w grupie która otrzymała kredyt są zbliżone do wartości w grupie która go nie otrzymała. Również możemy zauważyć wartości odstające, jednak nie są one aż tak wysokie i odbiegające od średniej jak w przypadku dochodów.



3. Metody k najbliższych sąsiadów

3.1 Przygotowanie danych

Dane zostały losowane podzielone na część uczącą i część testową. Część ucząca objęła 70% danych a testowa 30%.

Zmienną kategoryczną Property_Area zamieniono na zmienną ilościową zamieniając kategorię Urban/Semiurban/Rural na 0/1/2

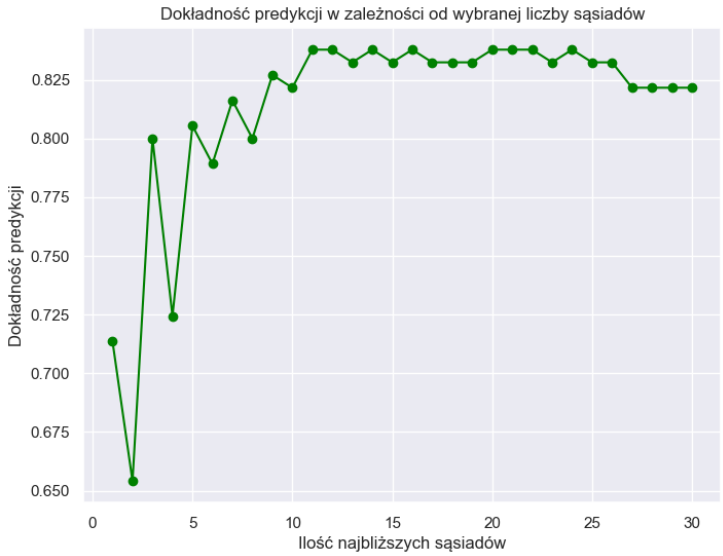
Następnie podzielone dane zostały znormalizowane w obu częściach. Normalizację wykonano dla zmiennych objaśniających które miały wartości inne niż 0/1. Były to Dependents, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Property_Area.

3.2 Metoda k najbliższych sąsiadów

Metoda k Najbliższych Sąsiadów (k-Nearest Neighbors) wyznacza k sąsiadów, do których badany element ma najbliżej dla wybranej metryki, a następnie wyznacza wynik w oparciu o większość wartości sąsiadów. Metoda ta może dawać znacznie różne wartości dla różnej liczby wybranych sąsiadów. Większe wartości k umożliwiają wygładzenie obszarów podziału, usunięcie szumu i artefaktów, lecz również prowadzą do błędów w klasyfikacji rzadszych wzorców.

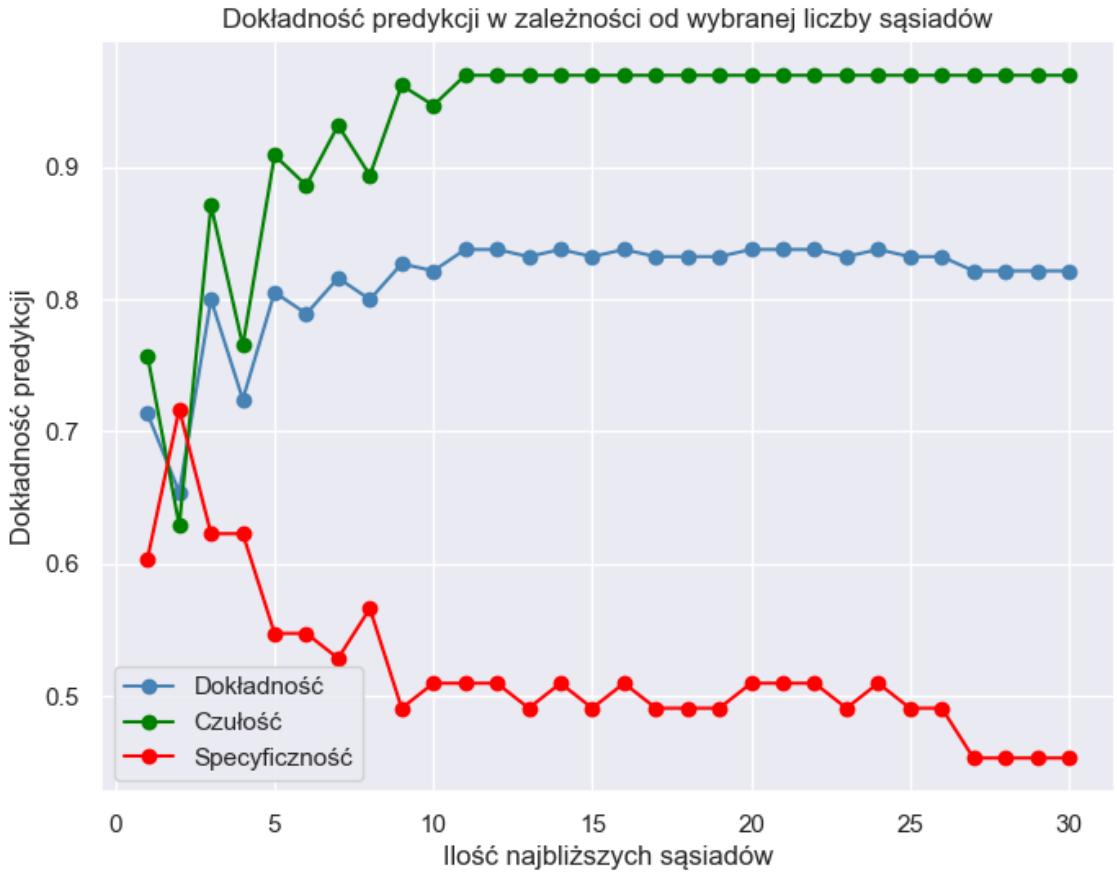
Dlatego też, aby dobrać odpowiednią liczbę rozpatrywanych sąsiadów możemy wykonać obliczenia dla wielu potencjalnych k i dla każdej z nich przygotować macierz błędów która pozwala nam obliczyć dokładność, czułość oraz specyficzność, gdzie dokładność oznacza stosunek poprawnych predykcji do ilości wszystkich wartości zbioru testowego, czułość oznacza stosunek poprawnych klasyfikacji 1 (prawda) dla wszystkich wartości równych 1 w zbiorze testowym, a specyficzność oznacza stosunek poprawnych klasyfikacji wartości 0 (fałsz) dla wszystkich wartości 0 w zbiorze testowym.

Używając metody k najbliższych sąsiadów byliśmy w stanie utworzyć poniższy wykres zależności liczby rozpatrywanych sąsiadów oraz dokładności.



Jednakże, dokładność nie zawsze jest odpowiednią miarą wyboru optymalnej liczby najbliższych sąsiadów dla danego problemu. W przypadku kredytów hipotecznych, bank może chcieć skupić się na uniknięciu ryzyka dania kredytu który nie zostanie spłacony. Dlatego też w danym przypadku możemy wybrać większą wartość specyficzności gdzie większa liczba wartości 0 (nieaktywny/nieotrzymany kredyt) jest sklasyfikowana poprawnie.

Porównanie ilości rozpatrywanych sąsiadów z dokładnością, czułością oraz specyficznością jest widoczna na poniższej grafice.



Jak możemy zauważyć, specyficzność spada wraz ze wzrostem czułości oraz dokładności.

W naszym przypadku dla optymalnej dokładności powinniśmy wybrać 14 jako liczbę najbliższych sąsiadów, jednak dla zmaksymalizowania specyficzności należy wybrać 2 jako liczbę sąsiadów. Pozwoli to zwiększyć specyficzność kosztem dokładności.

Poniżej zamieszczono macierze błędów dla danych uczących oraz testowych przy liczbie sąsiadów = 2

Loan_Status row_0	N	Y
N	139	56
Y	0	234

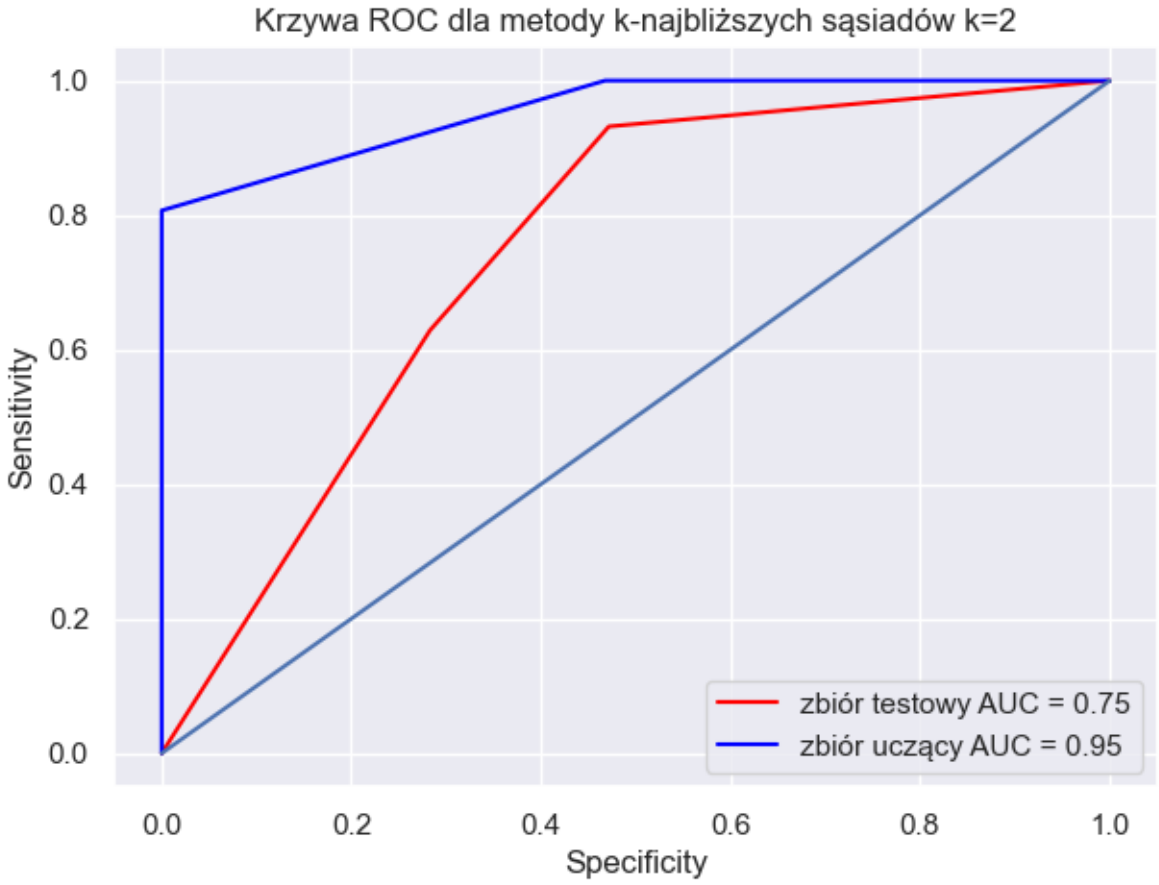
1Macierz błedów zbioru testowego

Loan_Status row_0	N	Y
N	38	49
Y	15	83

2Macierz błędów zbioru uczącego

Krzywa ROC

Krzywa ROC (Receiver Operating Characteristic) pozwala na ocenę jakości klasyfikacji badając zależność czułości i 1 - specyficzności. Obliczana jest ona dla poziomów prawdopodobieństwa klasyfikacji.



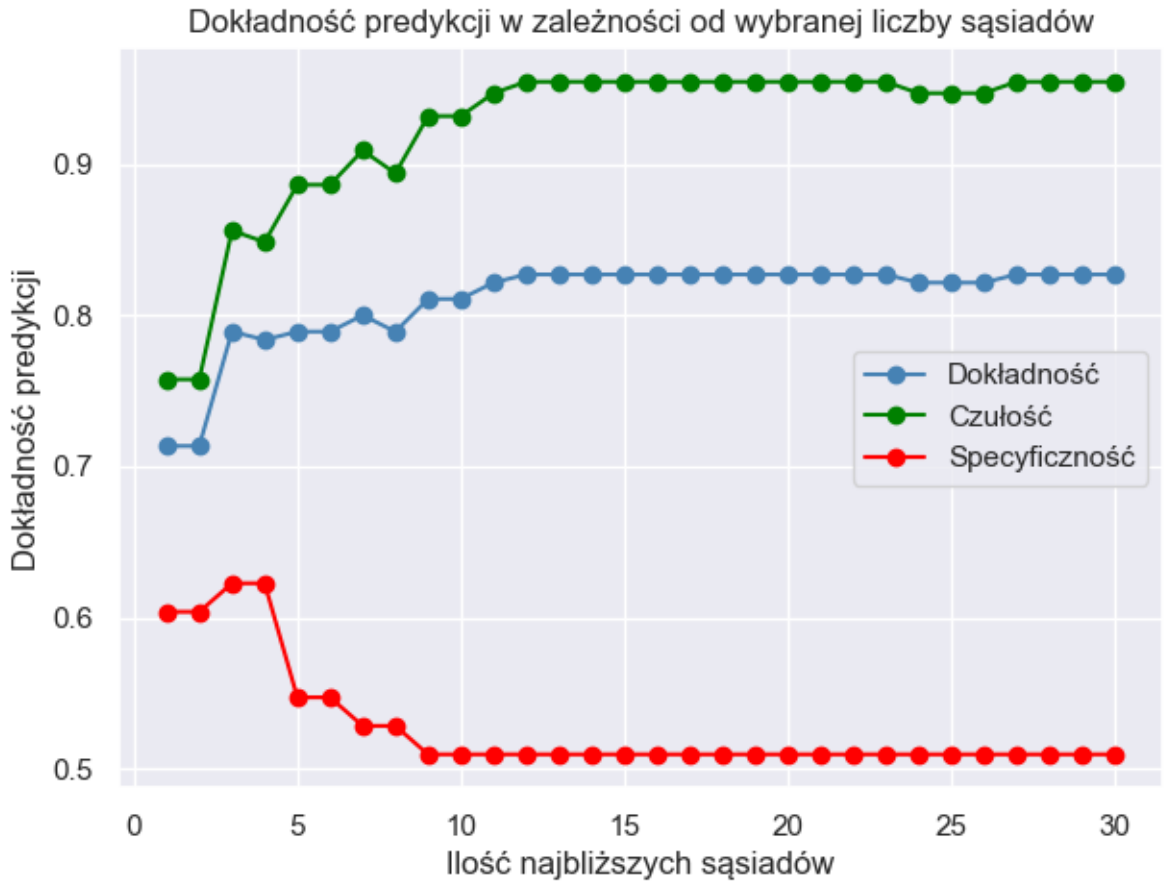
Dzięki krzywej ROC możemy zauważyć że wartości są lepsze niż teoretyczna linia klasyfikacji dla modelu losowego.

Wartość AUC pomaga określić poprawność klasyfikacji. Jest to pole pod krzywą. Im wartość jest bliższa 1, tym lepsza klasyfikacja. W badanych danych, AUC dla danych uczących wynosi prawie 1 (0.95) a dla zbioru testowego 0.75 co jest wystarczająco wysoką wartością aby uznać klasyfikację naszego modelu za poprawną.

3.3 Ważona metoda k najbliższych sąsiadów KKNN

Ważona metoda k najbliższych sąsiadów korzysta ze zbliżonej logiki co podstawowa metod najbliższych sąsiadów. Odróżnia ją korzystanie z wag dla każdego sąsiada. W badanym przypadku wagi określane są prez odległość od danego sąsiada. Sąsiedzi, którzy znajdują się bliżej poszukiwanego elementu mają większą wagę przy wyborze klasy.

Podobnie jak w poprzedniej metodzie, aby określić optymalną liczbę przyjętych sąsiadów, przygotowano wykresy ukazujące zależność dokładności, czułości oraz specyficzności od liczby wybranych sąsiadów.



Podobnie jak w przypadku poprzedniej metody specyficzność spada, a dokładność rośnie wraz ze wzrostem liczby wybieranych sąsiadów. Jednakże wartość specyficzności nie jest większa od dokładności w żadnym przypadku.

Optymalna liczba sąsiadów wynosi 3 patrząc pod kątem maksymalizacji specyficzności przy jednoczesnym zachowaniu wysokiej dokładności.

Co również ciekawe, w przypadku ważonej metody, klasyfikacja dla zbioru danych uczących osiąga dokładność na poziomie 100% czyli poprawnie klasyfikuje wszystkie wartości. Jednakże patrząc na niskie wartości specyficzności dla danych testowych, może to wskazywać na przetrenowanie modelu – 100% dokładności dla danych testowych nie przekłada się na lepsze wyniki dla danych testowych.

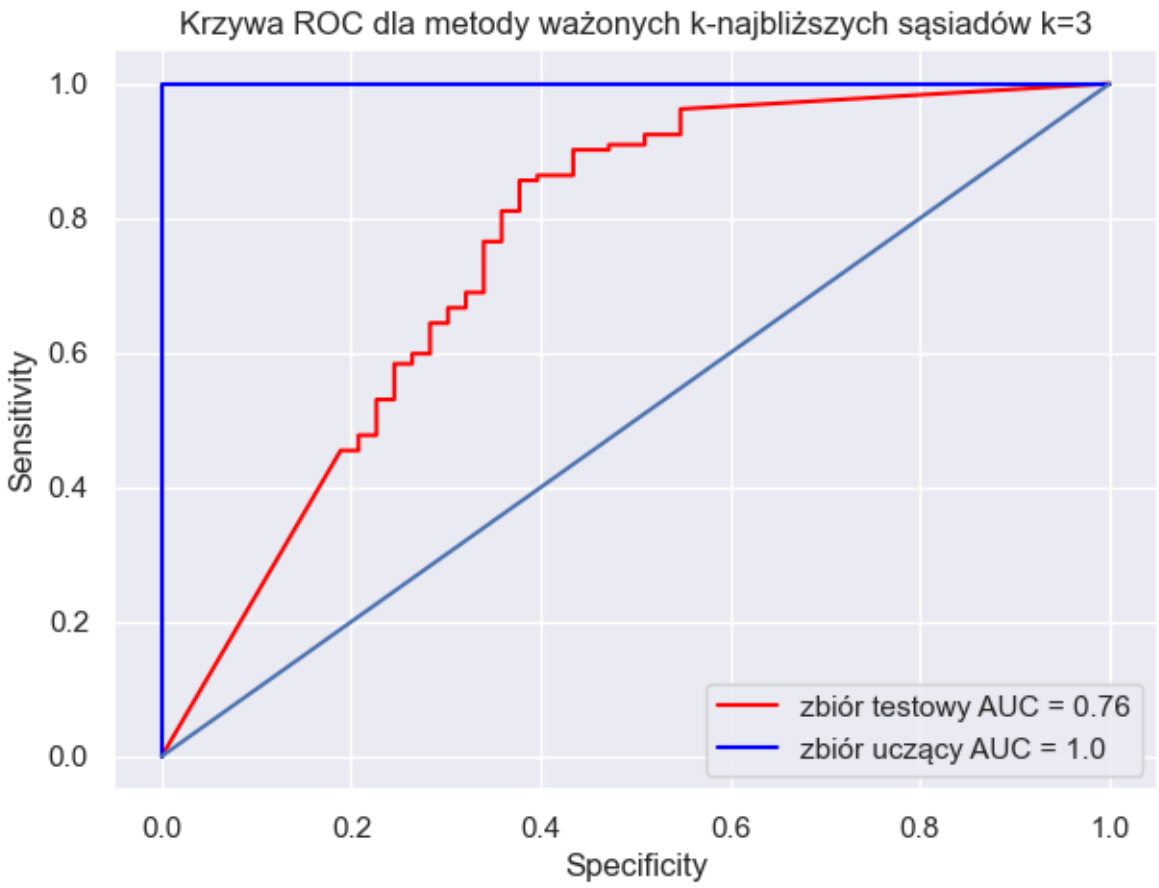
Loan_Status row_0	N	Y
N	139	0
Y	0	290

3Macierz błędów danych testowych dla k=3

Loan_Status row_0	N	Y
N	33	19
Y	20	113

4Macierz błędów danych uczących dla k=3

Krzywa ROC



Dla danych uczących wskaźnik AUC przyjmuje wartość 1 co pokazuje, że predykcje są całkowicie zgodne z danymi faktycznymi. Może to jednak wskazywać na przeuczenie modelu gdyż nie przekłada się to na wysokie wartości wskaźnika AUC dla danych testowych.

4. Klasyfikator naiwny Bayesa

Naiwny klasyfikator Bayesa – prosty klasyfikator probabilistyczny. Naiwne klasyfikatory bayesowskie są oparte na założeniu o wzajemnej niezależności predyktorów (zmiennych niezależnych). Do obliczenia prawdopodobieństwa przyjęcia danej klasy wykorzystuje twierdzenie teorii prawdopodobieństwa, wiążące prawdopodobieństwa warunkowe dwóch zdarzeń warunkujących się nawzajem, sformułowane przez Thomasa Bayesa.

4.1 Przygotowanie danych

Klasyfikator Bayesa wymaga danych kategorycznych do prawidłowego przeprowadzenia obliczeń.
Dane ApplicantIncome, CoapplicantIncome, LoanAmount zostały zamienione na klasy 0,1,2,3, które bazują na przedziałach percentyli danych zmiennych. Kolejno przedziałami były 0-percentyl(0.25)-mediana-percentyl(0.75)-maksymalna wartość.

W przypadku zmiennej Loan_Amount_Term utworzono 2 kategorie 0 dla okresu mniejszego niż 360 i 1 dla okresów równych lub większych niż 360. Wybrano podane granice przedziałów, ponieważ ponad 70% wartości wynosiło dokładnie 360.

Następnie wszystkie typ wszystkich kolumn zamieniono na dane kategoryczne.

Ponownie dane podzielono na część uczącą (70%) oraz testową (30%).

4.2 Predykcja z użyciem wytrenowanego modelu klasyfikacji naiwnej Bayesa

Model trenowano z użyciem danych uczących. Predykcja dla danych uczących pozwoliła na osiągnięcie 80% dokładności.

Dla danych testowych macierz błędów ukształtowała się jak widoczne na grafice poniżej.

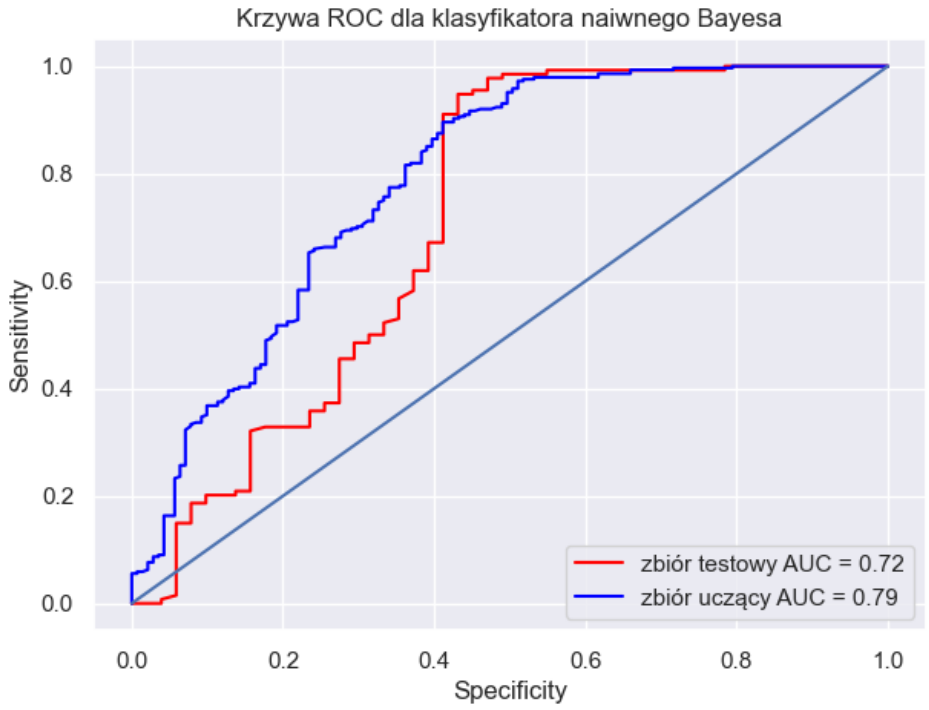
Loan_Status row_0	N	Y
N	25	2
Y	26	132

5Macierz błedów dla danych testowych

Na podstawie macierzy błędów obliczono wskaźniki klasyfikacji:

```
Dokładność: 0.8486486486486486
Czułość: 0.9850746268656716
Specyficzność: 0.49019607843137253
```

Jak można zauważyć, wartośc współczynnika czułości wyniosła niemal 1 ale specyficzność wyniosła poniżej 0.5 co jest bardzo niską wartością.



Wartości AUC obliczone na podstawie krzywej ROC są zbliżone dla zbioru uczącego i testowego. Krzywa dla zbioru testowego znajduje się w jednym miejscu poniżej linii potencjalnej klasyfikacji modelu losowego. Może to wynikać z bardzo niskiej specyficzności modelu.

Przyjmując potrzebę zminimalizowania specyficzności aby zmniejszyć ryzyko przyznania kredytu potencjalnie niewypłacalnemu kredytobiorcy, model bazujący na twierdzeniu Bayesa nie dałby wyników wystarczających dla przyjęcia klasyfikacji. Błąd w przypadku sklasyfikowania negatywnych odpowiedzi jest zbyt duży.

5. Podsumowanie

Zakładając chęć ograniczenia ryzyka i potrzebę zmaksymalizowanie specyficzności, najlepszym modelem klasyfikującym dane kredytowe byłby model oparty na metodzie k najmniejszych sąsiadów z k = 2. Ze wszystkich badanych modeli wygenerował klasyfikację o najmniejszej liczbie błędów przy klasyfikacji negatywnych odpowiedzi kredytowych. Przy tym charakteryzuje się zadowalającą wartością dokładności i AUC.

W przeciwieństwie do modelu opartego na ważonej metodzie k najbliższych sąsiadów nie jest modele przetrenowanym który generuje optymalne rozwiązanie jedynie dla danych uczących.

Model oparty na klasyfikacji naiwnej Bayesa dostarczył nierówne wyniki z dużą liczbą klasyfikacji false positive, gdzie prognozowane wartości były pozytywne dla realnie negatywnych wartości.
Jego relacja True negative – poprawnie sklasyfikowanych wartości negatywnych do wszystkicj rzeczywistych wartości negatywnych wyniosła mniej niż teoretyczny model losowy.