**PRACA DYPLOMOWA MAGISTERSKA**

MODELE REGRESJI W ZADANIACH KLASYFIKACJI DANYCH

**Wydział Fizyki Technicznej, Informatyki i Matematyki Stosowanej**

**Promotor: dr inż. Agnieszka Duraj**

**Dyplomant: Arkadiusz Ptak**

**Nr albumu: 186860**

**Kierunek: Informatyka**

**Specjalność: Systemy Informatyczne w Zarządzaniu i Gospodarce Elektronicznej**

Łódź (data)

Spis Treści

[1.1 Wybór tematu 3](#_Toc443991245)

[1.2 Cel pracy i zakres pracy 3](#_Toc443991246)

[2. Eksploracja danych 3](#_Toc443991247)

[2.1 Wstęp: 3](#_Toc443991248)

[2.2 Pojęcie eksploracji danych, oraz pochodzenie danych 4](#_Toc443991249)

[2.3 Co można eksplorować: 4](#_Toc443991250)

[2.4 Metody eksploracji danych: 5](#_Toc443991251)

[2.5 Problemy 6](#_Toc443991252)

[3. Klasyfikacja danych 7](#_Toc443991253)

[3.1 Metoda najbliższego sąsiada 8](#_Toc443991254)

[3.2 Naiwny klasyfikator bayesowski 9](#_Toc443991255)

[3.3 Drzewa klasyfikacyjne 11](#_Toc443991256)

[3.4 Lasy losowe 12](#_Toc443991257)

[3.5 Sieci neuronowe 12](#_Toc443991258)

[3.6 Metoda wektorów nośnych 14](#_Toc443991259)

[4. Regresja 15](#_Toc443991260)

[4.1 Pojęcie regresji 15](#_Toc443991261)

[4.2 Modele regresji 17](#_Toc443991262)

[4.2.1 Regresja liniowa 17](#_Toc443991263)

[4.2.2 Regresja Logistyczna 20](#_Toc443991264)

[4.2.3 Regresja wieloraka 21](#_Toc443991265)

[4.2.4 Regresja krokowa 25](#_Toc443991266)

[4.2.5 Regresja składowych głównych 26](#_Toc443991267)

[4.2.6 Regresja częściowych najmniejszych kwadratów 27](#_Toc443991268)

[4.2.8 Regresja nieliniowa 27](#_Toc443991269)

[4.2.9 Regresja nieparametryczna 28](#_Toc443991270)

[4.3 Zastosowanie regresji do klasyfikowania danych 29](#_Toc443991271)

[5. Implementacja 32](#_Toc443991272)

[6. Przeprowadzone badania 39](#_Toc443991273)

[6.1 Przedstawienie badań 39](#_Toc443991274)

[6.2 Analiza wyników 39](#_Toc443991275)

[7. Podsumowanie i wnioski 39](#_Toc443991276)

[8. Bibliografia 39](#_Toc443991277)

[9. Spis ilustracji 39](#_Toc443991278)

**1. Wstęp**

## 1.1 Wybór tematu

Do wyboru takiego tematu, głównie przyczyniła się wybrana przezemnie specjalizacja, jaką są Systemy Informatyczne w Zarządzaniu i Gospodarce Elektronicznej. W takich systemy przechowuja i przetwarzają dosyć sporą liczbę danych. Zatem, przy pracowaniu z tymi systemami bardzo istotną rzeczą jest przewidywanie i późniejsza klasyfikowanie danych. Według mnie dosyć interesującym narzędziem, które może łączyć te dwie rzeczy jest regresja, dlatego też wybrałem ją jako motyw przewodni poniższej pracy.

## 1.2 Cel pracy i zakres pracy

Celem tej pracy jest zapoznanie się z modelami regresji w zadaniach klasyfikacji danych. Czyli przetestowanie sposobu w jakim regresja może pełnić rolę klasyfikatora w dzieleniu danych na osobne zbiory. Poniżej w ramach wstępu opisana jest ogólnie eksploracja danych, następnie wspomniana jest klasyfikacja, oraz jej popularne metody. Kolejnym punktem jest regresja, gdzie opisane są jej główne modele, oraz zastosowanie regresji w klasyfikacji danych. Rozdział ten zamyka część teoretyczną, po której następuje część praktyczna, rozpoczynająca się rozdziałem zatytułowanym „Implementacja”. W tym rozdziale opisana jest aplikacja, która powstała na potrzeby badań, które przeprowadzono. Badania te przedstawione są w kolejnym punkcje „Przeprowadzone badania”. Do przeprowadzonych testów wykorzystano dwa modele regresji: model liniowy, oraz logistyczny. Są one popularne i świetnie nadają się do przedstawienia przykładów klasyfikacji. Na końcu znajduję się podsumowanie wraz z wnioskami, oraz biblioggrafia, spis tabel i rysunków.

# 2. Eksploracja danych

## 2.1 Wstęp:

Codziennie powstają nowe dane, które przetrzymywane są w wielkich ilościach. Ludzkie umiejętności analizowania i pojmowania tak ogromnych informacji są znacznie mniejsze niż możliwość ich przechowywania. Co z tego, że mamy mnóstwo danych, skoro nie potrafimy wydobyć odpowiedniej wiedzy, która jest w nich ukryta. Z problemem tym spotkało się wiele przedsiębiorstw oraz firm, dysponujących tak olbrzymimi zasobami danych. Ta sytuacja spowodowała zapotrzebowanie na technologie, która pozwala na efektywne wykorzystywanie zgromadzonych informacji i przekształcania ich na odpowiednią wiedzę. Dzięki temu swoje istnienie zapoczątkowała technologią zwana eksploracją danych. Zatem problem, którym dotychczas było magazynowanie wielkich zbiorów danych, zamienił się w atut, ponieważ z danych tych, można zaczerpnąć bardzo korzystne i użyteczne informacje.

## 2.2 Pojęcie eksploracji danych, oraz pochodzenie danych

W tym podrozdziale zawarta jest odpowiedź na pytania co to jest eksploracja danych i skąd biorą się ów dane

Zatem, co to jest eksploracja danych, a raczej jakie zadania ma przed sobą eksploracja lub też nazywana inaczej jako odkrywanie wiedzy w bazach danych?

Z definicji do zadań eksploracji danych należy automatyczne odkrywanie nietrywialnych dotychczas nieznanych, zależności, związków, podobieństw lub trendów (ogólnie nazywanych wzorcami) w dużych repozytoriach danych. Odkrywane w procesie eksploracji danych wzorce mają, najczęściej, postać reguł logicznych, klasyfikatorów (np. drzew decyzyjnych), zbiorów skupień, wykresów, itp.

Uogólniając celem eksploracji danych jest ich analiza co skutkuje lepszym poznaniem i zrozumieniem zgromadzonych informacji, tym samym stają się one bardziej użyteczne.

Zautomatyzowanie eksploracji danych daje nowe możliwości użytkownikowi w obszarze interakcji z bazą danych. Mianowicie pozwala na kreowanie zapytań przewyższających możliwości jakie daje nam popularny standard SQL.

Jak już wspomniano we wstępie eksploracja nie powstałaby bez tworzenia się ogromnej ilości danych. Skąd biorą się takie dane ? Odpowiedź jest dosyć oczywista, mianowicie stoi za tym informatyzacja praktycznie każdej dziedziny życia. Każdego dnia przez firmy, banki, sklepy, generowane są raporty, wykonywane są setki tysiące operacji itp. co skutkuje natłokiem ogromnej ilości danych. Kolejną dziedziną życia odpowiedzialną za olbrzymi przyrost danych jest nauka np. fizyka czy biologia, gdzie generuje się olbrzymie ilości danych powstałych podczas procesów eksperymentalnych. Również rozwój sieci internetowej przyczynił się do zwiększenia ilości generowanych danych. Powstało dużo stron internetowych, oraz rozwinął się tzw. e-handel.

## 2.3 Co można eksplorować:

W tym podrozdziale opisane zostanie co można eksplorować i jakie są metody tejże eksploracji.

Zatem, eksplorować możemy praktycznie każdy dowolny zbiór danych :

• Hurtownie danych

• Relacyjne bazy danych

• Repozytoria danych

• Przebiegi czasowe i temporalne bazy danych

• Przestrzenne bazy danych

• Tekstowe i multimedialne bazy danych

• Zaawansowane systemy informatyczne – Obiektowe i obiektowo-relacyjne bazy danych

Bardzo istotny jest wybór odpowiedniej metody eksploracji, która zostanie wykorzystana do analizy danego zbioru. Wymienione i opisane są one w następnym punkcie.

## 2.4 Metody eksploracji danych:

Eksploracji danych możemy dokonać według kilku metod podzielonych ze względu na cel eksploracji i typy wzorców odkrywanych w jej procesie

• odkrywanie asocjacji

• klasyfikacja/regresja

• grupowanie(klastrowanie)

• odkrywanie sekwencji

• odkrywanie charakterystyk

• wykrywanie zmian i odchyleń

• eksploracja tekstów

• eksploracja WWW

• analiza przebiegów czasowych

* Odkrywanie asocjacji: jest to najrozleglejsza klasa metod, która obejmuje metody odkrywania korelacji lub też interesujących zależności (nazywane są asocjacjami) w większych zbiorach danych. W wyniku działania odkrywania asocjacji powstają zbiory reguł asocjacyjnych bądź też wzorców sekwencji opisujących odkryte korelacje lub zależności. Bardzo dobrym przykładem takiej reguły jest reguła zakupu np. “Jeżeli klient kupił chleb i masło, zazwyczaj kupuje również ser”, w ten sposób możemy przewidzieć zachowanie klienta, co staje się bardzo użyteczne. Reguły takie można wykorzystywać do promocji produktów, bądź też poprawnego rozmieszczenia ich na pułkach.
* Klasyfikacja opisana jest ona dokładniej w następnym rozdziale.
* Grupowanie(klastrowanie, analiza skupień): polega na znajdowaniu skończonych zbiorów klas obiektów, które posiadają wspólne lub bardzo podobne cechy. Mamy kilka metod analizy skupień:
* Metody k-średnich, gdzie grupowanie polega na początkowym podzieleniu populacji na wcześniej ustaloną liczbę klas, a następnie niektóre elementy klas są transportowane do innych, tak aby uzyskać minimalna wariancję w uzyskanych klasach.

Metody hierarchiczne, gdzie dla zbioru obiektów tworzona jest hierarchia klasyfikacji, startując od momentu gdzie każdy obiekt stanowi jedno skupienie, a kończąc na takim podziale, gdzie wszystkie obiekty należą do tego samego skupienia.

Metody rozmytej analizy skupień, które są w stanie przydzielić poszczególny element do co najmniej jednej kategorii.

* Odkrywanie wzorców sekwencji: polega na analizowaniu bazy danych, która zawiera informacje o zdarzeniach występujących w określonym przedziale czasowym. Jej celem jest wykrycie powiązań między występowaniem konkretnych zdarzeń w czasie. Metoda ta ma zastosowanie w takich dziedzinach jak: telekomunikacja, medycyna, planowanie inwestycji giełdowych, przewidywania sprzedaży, analiza koszyków zakupów. Przykładem obrazującym tę metodę może być uczestnictwo w kursach: klient zamówił kurs dla początkujących, a za tydzień zamawia ten sam rodzaj kursu z tym, że dla średnio- zaawansowanych, a po upływie kolejnego - zamawia już kurs dla zaawansowanych. Podczas tych dwóch tygodni mógł on oczywiście uczęszczać też na inne kursy, ale istotne tutaj jest, to, że zamawiał on te same kursy z inną preferencją stopnia zaawansowania w regularnej sekwencji czasowej.
* Analiza przebiegów czasowych: polega na odnalezieniu pewnych trendów lub podobieństw występujących na przestrzeni przebiegów czasowych.
* Odkrywanie charakterystyk: polega na odnajdywaniu konkretnych zależności funkcyjnych występujących pomiędzy zmiennymi, które opisują zbiór danych.

metodę tę stosuje się np. w określaniu profilu klienta. Przykład zastosowania:

klienci kupujący krem na zmarszczki to głównie kobiety po trzydziestym roku życia itp. Kolejnym przykładem wykorzystania tej metody może być diagnozowanie objawów jakiejś choroby.

* Regresja: dokładniejszy opis regresji umieszczony jest w dalszym rozdziale, który jest jej całkowicie poświęcony.

## 2.5 Problemy

Nie wszystko jednak jest takie przyjemne na jakie wygląda.

Odkrywanie wiedzy w olbrzymich bazach, czy hurtowniach danych wiąże się z pewnymi problemami. Podczas eksploracji może zdarzyć się , że zostaną odkryte ogromne ilości reguł , których analizowanie pochłonie dużo czasu i będzie to całkowicie nieopłacalne.

Kolejnym problemem przed jakim staje eksploracja danych, jest zróżnicowanie wymagań poszczególnych użytkowników. Często użytkownicy są zainteresowani różnymi typami reguł z różnych relacji.

Operacje na tak złożonych i rozległych danych ciągnie za sobą duże problemy z efektywnością. Czasami odkrycie skomplikowanej metody, w dodatku na olbrzymiej ilości danych wiąże się z dużą złożonością obliczeniową, co skutkuje problemami z wydajnością.

# 3. Klasyfikacja danych

Ten rozdział poświęcony będzie dokładniejszym opisem pojęcia klasyfikacji danych.

Klasyfikacja danych to taka metoda analizy danych, która polega na określeniu wartości konkretnego atrybutu opierając się na posiadanym już zbiorze danych tzw, zbiór danych treningowych. Często metoda ta nazywana jest również jako: podział, grupowanie, analiza skupień, czy też identyfikacja. Wynikać to może z tego, że używana jest ona w wielu dziedzinach badawczych.

Poprzez sklasyfikowanie, możemy rozumieć przypisanie obiektów do predefiniowanych klas. Rozrózniamy dwa rodzaje klasyfikacji:

-Klasyfikacja dwuklasowa, w której obiekt przypisujemy do jednej z dwóch dostępnych klas (np. określamy czy dana cecha występuje, czy też nie).

-Klasyfikacja wieloklasowa, w tym przypadku obiekt możemy przypisać do więcej niż dwóch klas.

Klasyfikacja danych odbywa się w dwóch etapach. Pierwszym z nich jest etam tzw. uczenia się, w którym generowana jest „wiedza” używana pózniej do przypisywania obiektów do poszczególnych klas. Drugim etapem jest właśnie wspomniane przypiywanie (klasyfikowanie) obiektów do klas. W etapie pierwszym uczenie odbywa się z nadzorem eksperta – wprowadza on do modelu zarówno dane jak i klasyfikacje, na podstawie których w kolejnym kroku klasyfikowane są nowe obiekty ( z założenia dane wprowadzone przez eksperta, są prawidłowe).

Zatem klasyfikacje danych możemy zakwalifikować jako metoda nadzorowana.

W metodzie klasyfikacji dane najczęściej spotyka się w postaci systemów informacyjnych gdzie reprezentowane są za pomocą tabel, np.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Badany | Mdłości | Zawroty | Gorączka | Przeziębiony |
| 1 | NIE | TAK | TAK | TAK |
| 2 | TAK | NIE | NIE | NIE |
| 3 | TAK | TAK | TAK | TAK |
| 4 | NIE | TAK | NIE | NIE |

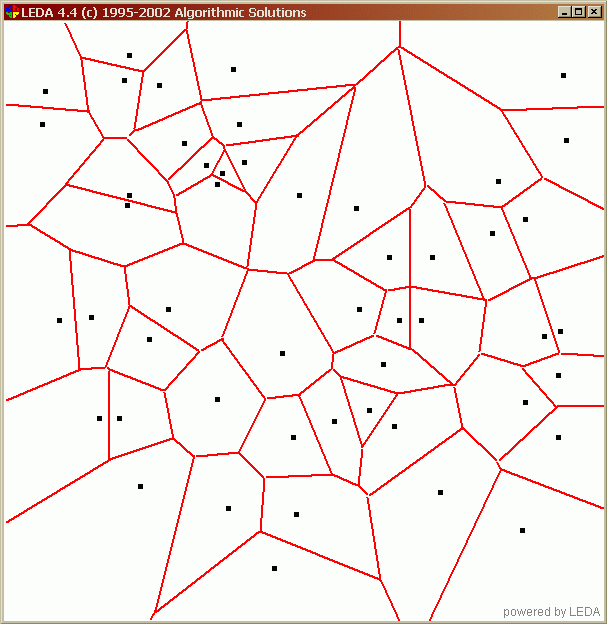
Klasyfikacja jest wykorzystywana przy rozpoznawaniu obiektów w dużych bazach danych zawierających obrazy, czy też w medycynie przy procesie klasyfikacji pacjentów do poszczególnych schorzeń.

Rozróżniamy kilka metod klasyfikujących dane:

### 3.1 Metoda najbliższego sąsiada

Chyba najbardziej popularna metoda klasyfikacji Polega ona na tym, że podczas rozważania do której klasy przypisać dany obiekt, bieże się pod uwagę klasę obiektu sąsiedniego, czyli takiego, który jest w najbliższej odległości od badanego i umieszcza się w niej nasz nowy obiekt. Odległość ta wyliczana jest przy pomocy konkretnych funkcji, których wybór jest spory.

Klasyfikacja przy pomocy tej metody zaliczana jest do rzędu nieparametrycznych, co za tym idzie, nie wymaga ona jakichkolwiek założen co do rozkładów danych w klasach. Do zobrazowania tej metody, świetnie nadaje się diagram Woronoja, który dokonuje podziału jednej płaszczyzny na kilka rozłacznych komórek, posiadających zaledwie jeden punkt danych wejściowych. Krawędzie tych komórek są jednakowo odległe od dwóch sąsiadujących z nimi punktów(linia dzieląca te dwa punkty znajduje się dokładnie w połowie).



Rysunek 1 Diagram woronoja

Gdy liczba prób rośnie nieskończenie metoda ta jest bardzo efektywna, jednak w praktyce, dużo jest przypadków gdzie nie dysponujemy aż tak wielką liczbą obserwacji.

Zaletą tej metody, jest również jej prosta implementacja, co spowodowane jest tym, że nie wymaga ona estmacji warunkowych funkcji gęstości. Wartość parametru K – który określa nam liczbę sąsiadów, dobierana jest w sposób eksperymentalny, przy pomocy próby testowej, lub stosując metodę sprawdzania krzyżowego. Aby uniknąć losowości, która może wystąpić ze względu na błedne sklasyfikowanie cechy stosujemy normalizację, która zapobiega „przyćmienie” cech o mniejszych wartościach, przez te z wartościami większymi.

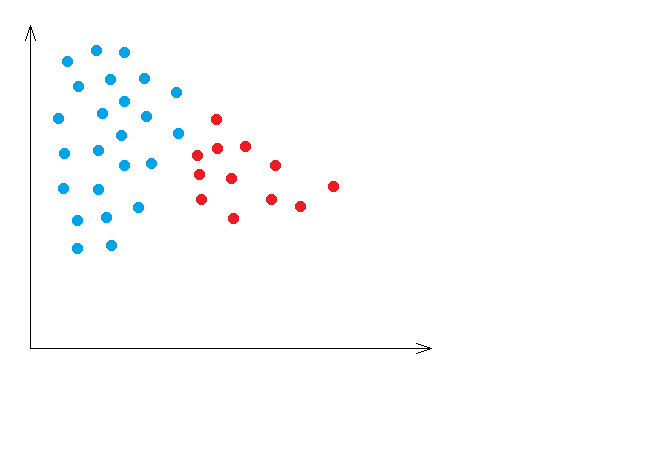
Zatem dla zmiennych ciągłych korzystamy z normalizacji min-max:

Zastosowanie normalizacji jest konieczne, gdyż przekształca ona zmienne o różnych wielkosciach do postaci porównywalnej.

Kolejnym pozytywnym aspektem metody najbliższego sąsiada jest jej odporność na wszelkie zakłócenia i odchyły. Kosztem tego jest jednak duża złożoność obliczeniowa, rosnąca proporcjonalnie do liczby prób. Nie możemy także w żaden sposób wesprzeć się zmiennymi jakościowymi, a rezultat takiej klasycikacji jest bardzo mocno zależny od dostosowanej przez nas miary odlegóści.

### 3.2 Naiwny klasyfikator bayesowski

Najwny klasyfikator bayesowski jest prostym klasyfikatorem, którego założeniem jest, wzajemna niezależność pomiędzy cechami, które opisują klasyfikowane obiekty. Słowo naiwny w nazwie, nie wzieło się znikąd, metodę tę nazwano w ten sposób, ponieważ założenie jakim jest wspomniana niezależność, praktycznie nie występuje. Mimo tego, klasyfoikator ten radzi sobie całkiem nieźle ze swoim zadaniem, czyli grupowaniem obiektów, a w szczególności gdy jest ich spora liczba. Przyjżyjmy się prostemu przykładowi, który zobrazuje metodę działania klasyfikatora:



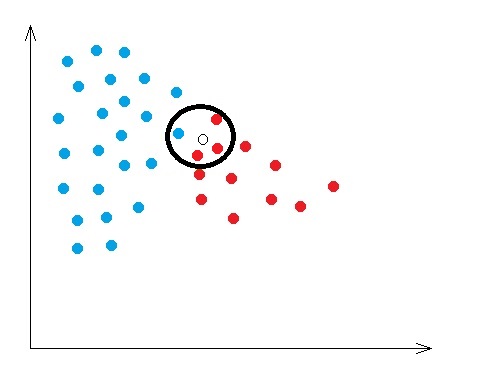
Rysunek 2 Działanie klasyfikatora bayesowskiego

Jak widzimy na powyższym rysunku, mamy zbiór punktów z podziałem na klasy niebieskie i czerowne. Przed nami stoi zadanie zaklasyfikowania kolejnego punktu.

Jak można zauważyć, niebieskich punktów jest dwukrotnie więcej niż czerwonych , zatem logicznym stwierdzeniem będzie, że nowo przyjęty punkt, powinien być sklasyfikowany jako niebieski (ze względu na prawdopodobieństwo). Takie prawdopodobieństwo nosi nazwę prawdopodobieństwa a priori, polegają one na posiadanych obserwacjach (a priori). W naszym przypadku mowa jest o procent punktów niebieskich względem punktów czerwonych. A priori są użyteczne przy przypisywaniu do klas nieznanych przypadków, zanim się pojawią.

Możemy więc przyjąć, że prawdopodobieństwa a priori (APR)dla klas wyliczamy następująco:

Mając wyliczone te prawdopodobieństwa, przystąpić możemy do sklasyfikowania nowego obiektu. Prawdopodobieństwo a priori przemawia za tym, że nowy punkt sklasyfikowany będzie do grupy niebieskiej. Należy wziąć również pod uwagę położenie klasyfkowanego punktu. Dodajemy zatem okrąg, obejmujący jego najbliższych sąsiadów, bez względu do jakiej klasy należą. W ten sposób będziemy mogli obliczyć wartość zwaną szansą.



Rysunek 3 Klasyfikator bayesowski biorący pod uwagę liczbę sąsiadów

W odniesieniu do powyższej sytuacji, szansa() na to, że nowy obiekt sklasyfikowany zostanie jako niebieski, czy czerwony obliczamy:

gdzie :

- liczba obiektów danego koloru w obrębie punktu klasyfikowanego

- całkowita liczba obiektów danego koloru

W takim przypadku dla naszej sytuacji, bardziej prawdopodobne jest, że obiekt zostanie przypisany do grupy czerwonej, gdyż obiekty tego koloru w sąsiedztwie są w przewadze.

Przystępujemy do końcowej klasyfikacji wedlug reguły Bayesa:

gdzie k to rozważany kolor. Wybieramy tę klasę dla której współczynnik C będzie wyższy, w naszym przypadku jest to klasa kolorów czerwonych.

### 3.3 Drzewa klasyfikacyjne

Drzew klasyfikacyjnych używamy kiedy zmienna zależna wyrażona jest na skali porządkowej lub też nominalnej.

Pod pojęciem drzewa kryje się graficzny model, który powstał po rekurencyjnym podziale zbioru obserwacji A na n rozłącznych zbiorów . Model taki budujemy w celu pozyskania maksymalnie jednorodnych podzbiorów z punktu widzenia wartości zmiennej zależnej. Proces taki skłąda się z wielu etapów, a w każdym z nich może wykorzystywać kompletnie inną zmienną niezależną. Podczas każdego etapu następuje analiza wszystkic predyktorów, z których wybierany jest jeden zapewniający najlepsze podział węzła (zawierający maksymalnie jednorodne podzbiory). Początkowy zbiór, który stanowi nasze całe drzewo, poddawany jest podziałowi na conajmniej dwa podzbiory. Jeżeli podział ograniczy się tylko do dwóch podzbiorów, drzewo takie nazywamy binarnym. W przeciwnym wypadku mamy do czynienia z drzewami dowolnymi. Zbiór, który dzieliliśmy na mniejsze określamy jako węzeł macierzysty, a podzbiory, które powstały w wyniku podziału otrzymują miano wezłów potomków. Następnie taki węzeł potomków pełni rolę węzła macierzystego a węzeł które nie zostały podzielone, stają się węzłami końcowymi (liśćmi). Przy pomocy liści określamy wielkość drzewa (ilość lisci=wielkość drzewa), jego głebokość stanowi liczba krawędzi pomiędzy pierwszym węzłem macierzystym (tzw. wierzchołkiem) a najodleglejszym liściem. Istnieją 3 główne grupy z których wywodzi się znaczna większość algorytmów drzew klasyfikacyjnych:

-CLS: algorytm ten stosowany był do binarnego dzielenia zbiorów obiekietów na dwa rodzaje klas: pozytywną (K+) oraz negatywną (K-). Wybiera on wartość, która najbardziej różniła obiekty należące do tych dwóch klas.

-AID: powstał jako alternatywne rozwiązanie dla regresji. Jego rozwinięciem jest algorytm WAID, ale najpopularniejszą odmianą tego alogyrtmu jest CHAID, w którym, na każdym etapie dzielącym drzewo, tworzy się tabelę kontygencji – zestawiającą zmienną niezależną z zależną. Algorytm ten umożliwia budowanie modeli dyskryminacyjnych, w których predykowana zmienna jest nominalna.

-CART: jest jedną z bardziej rozbudowanych metod slużących do budowania drzew klasyfikacyjnych. W algorytmie tym, do podziału drzewa stosowane są takie metody jak: miara netropii, reguła podziału na dwie części, czy też indeks Giniego. Używany tutaj proces rozrostu drzewa i przycinania jego gałęzi dla redukcji opisu liści , jest charakterystyczny dla tej metody. Przycinanie liści odbywa się przy nieznacznym wzroście błedu klasyfikacji, co pozwala na porównanie drzewa rozbudowanego oraz tego z obciętymi węzłami.

### 3.4 Lasy losowe

Lasy losowe są metodą, w której łaczymy ze sobą dużą ilość drzew klasyfikacyjnych i zaliczane są do procedur agregujących. Na początku realizowania tej metody, losujemy pewną liczbę prób bootstrapowych (próba powstała przez losowanie ze zwracaniem n elementów z X). Dla każdej takiej próby tworzymy drzewo klasyfikacyjne(bez przycinania), opierając się na tym, że w każdym węźle musi być wylosowana liczba cech będąca mniejszą niż ogólna liczba wszystkich cech. wylosowane cechy będą brały udział w wybieraniu najlepszego podziału. Końcowa klasyfikacja odbywa się przy pomocy głosowania większościowego nad klasami, które wskazane były przez drzewa decyzyjne.

Lasy losowe, charakteryzują się tym, że tworzy się je w dość łątwy i szybki sposób, zatem wykorzystywać je można nawet do klasyfikacji dużych zbiorów danych. Mogą one być równierz wykorzystywane w analizie przeżycia lub regresji.

### 3.5 Sieci neuronowe

Kolejnym popularnym modelem klasyfikującym dane, są sieci neuronowe. Jej głównym składnikiem jest tzw. neuron, który składa się z wartości wejściowych , wag odpowiadającym tym wartością, bloku sumacyjnego, funkcji aktywacji oraz sygnału wyjściowego . Na wejściach mamy dane, które doprowadzane są z neuronów warstwy poprzedniej, oraz zazwyczaj wartość bias – dodatkowe wejście, którego wartość jest stałą. Następnie każda wartość sygnału początkowego mnożona jest przez odpowiadającą mu wagę. Możemy w ten sposób manipulować sygnałami wejściowymi i decydować o ich wpływie na końcową wartość sumy. Wartość wagi może być zarówno dodatnia jak i ujemna. Kolejnym krokiem jest zsumowanie wartości otrzymanych po wymnożeniu sygnałów wejściowych przez odpowiadające im wagi i przekazanie tej sumy jako argument do funkcji aktywacji.

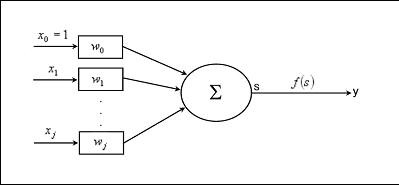
Matematyczny wzór neuronu prezentuje sięnastępująco:

gdzie:

f – to funkcja aktywacji

- to wartości sygnałów wejściowych, wraz z wartościami wag im odpowiadających

- Waga dla wartości wejściowej bias



Rysunek 4 Neuron

Funkcja aktywacji, może przyjmować postać liniową,nieliniową lub skoku jednostkowego(funkcja progowa). Jej wybór zależy od rodzaju problemy jaki dana sieć ma rozwiązać. Funkcja taka powinna spełniać następujące warunki:

- zapewnienie ciągłości przejścia między własną wartością minimalna a maksymalna

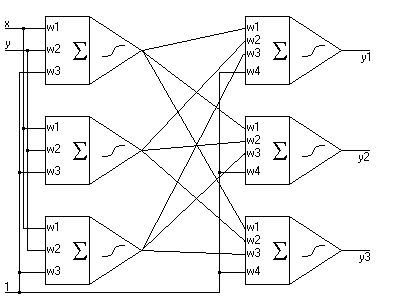
- łatwość jej obliczania i ciągłość pochodnej

- możliwość wprowadzenia parametru beta do ustalenia kształtu krzywej.

Do najczęsciej stosowanych funkcji należą: sigmoidalna(krzywa logistyczna), która przyjmuje wartości od 0 do 1, a także tanges hiperboliczny o wartośćiach z przedziału -1 do 1.

Funkcje nieliniowe wykorzystywane są często w sieciach wielowarstowych, ponieważ neurony w tych sieciach charakteryzują się dużą zdolnością do nauki.

Sieć wielowarstwowa, to nic innego jak połączone ze sobą neurony, w taki sposób, że wartość wyjściowa jednego jest jednocześnie sygnałem wejściowym do kolejnego.



Rysunek 5 Wielowarstwowa sieć neuronowa

Istnieją dwa typy architektury sieci: sieć jednokierunkowa, gdzie mamy jeden kierunek przepływu sygnału oraz sieć rekurencyjna, w której występuje zjawisko sprzężenia zwrotnego.

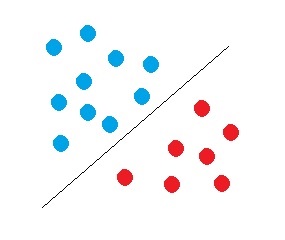
Aby sieć neuronowa działała poprawnie, musimy ją najpierw „nauczyć”. Proces nauki odbywać się może na dwa sposoby:

- Uczenie nadzorowane, gdzie korzystamy ze zbioru uczącego, w którym zawierają się pary sygnałów wejściowych wraz z wartością wyjściową. Sieć ma na celu stworzenie jak najlepszej funkcji przybliżającej powiązania między wejściem a wyjściem. Różnica jaką otrzymamy pomiędzy wyliczoną wartościa wyjściową, a pożądaną jest naszą miarąbłedu, która wykorzystywana jest przy korekcji wag. Typową metodą uczenia sieci wielowarstwowej jest wsteczna propagacja, gdzie każdy neuron zmniejsza swój błąd stosując metodę spadku gradientu.

- Uczenie nienadzorowane, które stosujemy, gdy nie znamy wartości oczekiwanych. warunkiem tego uczenia jest, aby sygnały wejściowe dały się w jakiś sposób sklasyfikować oraz liczba neuronów musi być wystarczająco duża, aby „zaabsorbować” przekazaną jej wiedzę.

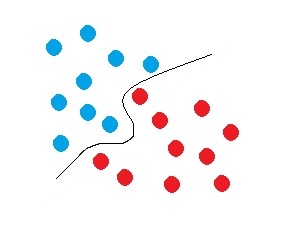
### 3.6 Metoda wektorów nośnych

Metoda wektorów nośnych polega na odseparowaniu obiektów o róznej przynależności klasowej. Proces ten zachodzi poprzez tworzenie przestrzeni hiperpłaszczyzny, które oddzielają obiekty należące do różnych klas. Prosty przykład takiej klasyfikacji zaobserwować możemy na poniższym rysunku:



Rysunek 6 Metoda wektorów nośnych - liniowy podział

Przykład ten jednak jest bardzo prosty, ponieważ mamy do czynienia z sytuacją idealną, gdy klasyfikacji możemy dokonać liniowo. Zazwyczaj jednak wygląda to nieco bardziej skomplikowanie:



Rysunek 7 Metoda wektorów nośnych - nieliniowy podział

Każda zmienna skategoryzowana otrzymuje zestaw zmiennych określających przynależność każdego przypadku. Gdy mamy zmienną która przyjmuje wartości A,B i C, będzie posiadała reprezentację w postaci: .

Hiperpłaszczyznę, która będzie jak najbardziej optymalna tworzymy za pomocą iteracyjnego algorytmu uczącego/ który za zadanie ma znimimalizować funkcję błedu.

# 4. Regresja

W poniższym rozdziale szczegółowo przyjrzymy się zagadnieniu regresji,

poznamy jej definicje, rodzaje oraz zastosowanie.

## 4.1 Pojęcie regresji

Regresja to nic innego jak wyrażenie pewnej zależności pomiędzy wartością jednej zmiennej od wartości drugiej zmiennej, przy użyciu nieskomplikowanej funkcji. W efekcie, na podstawie tej zależności, jesteśmy w stanie wyestymować wartość interesującej nas zmiennej na podstawie innej.

Wartość, którą próbujemy przewidzieć, nazywamy zmienna objaśnianą, a zmienne na podstawie których to robimy, noszą nazwę zmiennych objaśniających. Wartości zmiennych objaśniających są w ustalane przez osobę przeprowadzającą eksperyment, co za tym idzie, są przez nią całkowicie kontrolowane - skutkuje to pozbawieniem ich elementu losowości. Regresję w której występuje więcej niż jedna zmienna objaśniana, nazywamy regresją wieloraką, ten model jest opisany w następnym podrozdziale.

Praktyczne zastosowanie regresji ogranicza się do dwóch etapów:

* Pierwszym z nich jest budowanie modelu regresyjnego (tzw. konstruowanie modelu). Polega na utworzeniu funkcji, która opisuje zależność wartości oczekiwanej zmiennej objaśnianej od wartości zmiennej/zmiennych objaśniających. Funkcja może być podana w postaci wzoru matematycznego, ale nie tylko. Często stosuje się przypadki, gdzie rolę funkcji pełni konkretny algorytm, np. drzewo regresyjne albo sieć neuronowa. Model regresyjny tworzony jest w taki sposób, aby był jak najlepiej dopasowany do danych, które wprowadzamy. Dane te zawierają zestaw zmiennych objaśnianych jak i objaśniających. Zbiór takich danych nazywamy zbiorem uczącym, ponieważ po jego przeanalizowaniu nasza funkcja, lub algorytm potrafi przewidzieć wartości kolejnych zmiennych. Aby wyniki były zadowalające taki zbiór musi być odpowiednio duży. Oprócz zbioru uczącego powinniśmy posiadać również zbiór testowy. Potrzebny jest on do weryfikacji naszej funkcji czy też algorytmu. Taki zbiór składa się również ze zmiennych objaśniających i odpowiadającej im zmiennej objaśnianej. Przyjmuje się, że zbiór testowy powinien stanowić ⅓ zbioru uczącego. Na tym właśnie polega wyliczanie regresji.
* Gdy mamy zbudowany model, możemy już go stosować. Stosowanie modelu - bo tak nazywa się drugi etap, polega na zastosowaniu wyliczonego modelu przy użyciu danych, gdzie znane są tylko zmienne objaśniające. Jeżeli model został utworzony prawidłowo, wynikiem końcowym tej operacji są pożądane przez nas zmienne objaśniane.

Ogólna postać modelu wygląda następująco:

gdzie:

x - wektor zmiennych objaśniających

Y- zmienna objaśniana

- wektor współczynników regresji

f(x,) - funkcja regresji o wartościach w liczbach rzeczywistych

- błąd losowy o rozkładzie losowym

Odkrywcą zjawiska regresji do średniej był Sir Francis Galton. Była to dosyć ciekawa teoria, która wywołała oburzenie wyznawców teorii ewolucji. W skrócie, teoria ta polegała na tym, że ze związku dwóch wysokich rodziców powstawał potomek od nich niższy (dążący do średniej), a ze związku dwójki niskich rodziców rodził się potomek od nich wyższy (także dążący do średniej). Teorię tę rozwijał później Karl Pearson.

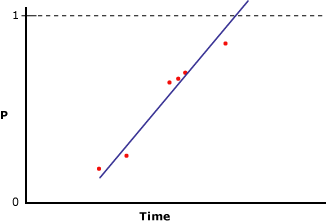
Działem statystyki, który zajmuje się metodami i modelami regresji jest analiza regresji.

## 4.2 Modele regresji

### 4.2.1 Regresja liniowa

Regresja liniowa jest jedna z najpopularniejszych regresji. W tym modelu, funkcja przyjmuje postać liniową:

gdzie y jest szacowaną wartością zmiennej celu, jest nachyleniem linii regresji lub współczynnikiem kierunkowym, a jest punktem przecięcia linii regresji z osią y (wyrazem wolnym) czyli wartością zmiennej objaśnianej w przypadku gdy zmienna objaśniająca jest równa zeru (w wielu przypadkach taka wartość zmiennej objaśniającej nie ma większego sensu).. Zmienne i są nazywane współczynnikami regresji.



Rysunek 8 Przykładowy wykres regresji liniowej

Stosując wzór:

możemy napotkać na sytuację, gdy w naszej bazie uczącej będzie taka para zmiennych objaśnianych, która ma rożna wartość, a daje taką samą zmienną objaśniającą (np. przy estymowaniu średniego spalania paliwa przez samochody, bierzemy pod uwagę liczbę koni mechanicznych, okazuje się, że pojazd mający moc 100KM osiąga takie samo spalanie jak ten posiadający 120KM). Dlatego tak naprawdę równanie, które reprezentuje prawdziwą liniową zależność ma postać :

pozwala nam to wyeliminować zależność zmiennej objaśniającej od jakości próby, co umożliwia nam stosowanie modelu, do wprowadzania danych spoza próby.

Wartość to tzw. losowy błąd, który stosowany jest w celu uwzględnienia nieoznaczoności w modelu. Równanie to nazywamy prostą regresji lub też równaniem regresji.

Linia prosta uzyskana przy pomocy funkcji, używana jest do przybliżenia relacji pomiędzy pojedynczą zmienną objaśniającą, a pojedynczą zmienna objaśnianą. Linia ta w modelu regresji liniowej wyznaczana jest za pomocą metody najmniejszych kwadratów. Jest to jedna z najważniejszych i najdłużej używanych w statystyce metod obliczeniowych. Jej celem jest wyznaczenie interesującej nas linii regresji, dla wprowadzonych danych. Można wykorzystywać ją do szacunku zależności liniowych jak i nieliniowych. Aby metodą najmniejszych kwadratów poprawnie wyliczyć współczynniki regresji, należy skorzystać z następujących wzorów:

Suma kwadratów błędów dla populacji:

=

Gdzie są wartościami doświadczalnymi, pobranymi ze zbioru uczącego, a n to liczba wykonywanych pomiarów.

Jeżeli wyznaczyliśmy już wzór funkcji regresji i chcemy wyznaczyć wartość zmiennej objaśnianej dla nowych zmiennych objaśniających, to będzie leżeć ona dokładnie na linii funkcji regresji. Natomiast zdarzać się będzie, że po sporządzeniu wykresu i podstawieniu danych ze zbioru współrzędne punktu znajdować się będą nie bezpośrednio na linii regresji, lecz pod lub nad nią. Załóżmy, że w zbiorze uczącym wartość oczekiwana była równa 100, natomiast po podstawieniu do wzoru, otrzymaliśmy 105, wtedy różnica pomiędzy tymi wartościami (w tym przypadku 5) będzie naszym błędem predykcji. Do minimalizacji tego błędu potrzebna jest nam wspomniana wyżej metoda najmniejszych kwadratów.

Model liniowy możemy stosować, kiedy spełnione są następujące warunki:

* zmienne objaśniające są liniowo niezależne
* zmienne są niezależne o rozkładzie N(0, σ2 ),
* wartość oczekiwana Y wyraża się przez liniową kombinację zmiennych X

Nigdy nie ma pewności, że linia regresji wyznaczona powyżej opisanym sposobem jest w stu procentach skuteczna, Istnieje jednak pewien sposób, aby określić przydatność takiej funkcji. Do określenia ów przydatności możemy posłużyć się współczynnikiem determinacji Szacuje on w jakim stopniu linia regresji najmniejszych kwadratów określa zmienność obserwowanych danych. Aby dokonać oceny takiej przydatności musimy obliczyć sumę kwadratów błędów oszacowania, która określa nam całkowitą wartość błędu przy zastosowaniu regresji liniowej.

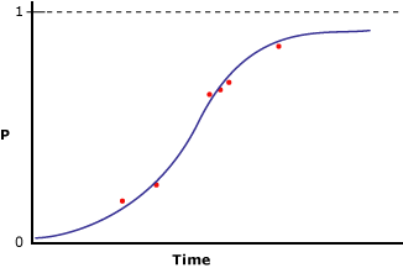
Do obliczenia tej sumy używamy wzoru:

gdzie to wartość zmiennej objaśnianej z próby, a to wartość zmiennej objaśnianej wyliczona z równania regresji.

Następnie taką samą czynność powtarzamy dla wartości uśrednionych. Nie bierzemy pod uwagę zmiennych objaśnianych, a wartość wyznaczamy za pomocą wyliczenia średniej ze wszystkich wartości . Obliczyliśmy w ten sposób całkowitą sumę kwadratów, którą możemy przyrównać do wyliczonej wcześniej sumy kwadratów błędów oszacowania. W ten sposób przekonamy się jak bardzo poprawiła się nasza estymacja, jeżeli skorzystaliśmy z modelu regresji liniowej.

### 4.2.2 Regresja Logistyczna

Model ten stosowany jest w przypadkach, w których nie mamy pewności, że zmienna objaśniana jest w postaci ciągłej. W takiej sytuacji mamy do czynienia ze zmienną jakościową, więc model regresji liniowej nie spełni naszych oczekiwań. Regresja logistyczna jest pod wieloma względami podobna do regresji liniowej, więc jest jedną z metod opisu zależności pomiędzy zmienną objaśnianą, a zmiennymi objaśniającymi. Charakterystyczną cecha tego modelu jest to, że oczekiwana przez nas zmienna objaśniana jest zmienną binarną - tzn. że przyjmuje wartości 0 lub 1 (zdarzenie występuje, lub nie występuje). Kolejną charakterystyczną cechą dla regresji logistycznej, jest jej wykres, który występuje w postaci krzywych sinusoidalnych, układających się w kształt litery “S”.



Rysunek 9 Przykładowy wykres regresji logistycznej

Rozważając zmienną objaśniającą Y przy danej wartości X=x oznaczoną jako E(Y|x), stosując liniowy model regresji, mielibyśmy wzór wyglądający następująco:

dla regresji logistycznej wzór wygląda nieco inaczej:

Przy stosowaniu modelu regresji logistycznej, użytecznym przekształceniem, jest funkcja logitowa, która opisana jest następującym wzorem:

funkcja ta posiada kilka przydatnych cech modelu regresji liniowej, takich jak: zakres od minus nieskończoności do plus nieskończoności, ciągłość oraz liniowość.

W regresji liniowej, mogliśmy skorzystać z metody najmniejszych kwadratów, natomiast w modelu logistycznym, metoda ta nie ma zastosowania, więc musimy korzystać z nieco innego rozwiązania.

Z pomocą przychodzi metoda estymacji największej wiarygodności, która umożliwia nam wyznaczenie parametrów o największej wiarygodności obserwowanych danych.

Funkcja wiarygodności jest to funkcja parametrów , która określa prawdopodobieństwo uzyskania obserwowanych danych x. Znajdując wartości , które maksymalizują funkcje wiarygodności odkrywamy estymatory największej wiarygodności - są to najbardziej wiarygodne wartości parametrów dla obserwowanych danych.

Funkcja wiarygodności prezentuje się następująco:

L()=

gdzie to prawdopodobieństwo wystąpienia zdarzenia, aprawdopodobieństwo nie wystąpienia zdarzenia.

Estymatory największej wiarygodności wyznaczamy poprzez różniczkowanie funkcji L() względem każdego parametru, a następnie przyrównywując je do zera.

### 4.2.3 Regresja wieloraka

W poprzednich modelach głównym założeniem było to, że zmienna objaśniana jest zależna tylko od jednej zmiennej objaśniającej. Istnieją jednak przypadki, gdzie oczekiwany przez nas wynik zależy od więcej niż jednego parametru. Do rozwiązywania takich problemów wymyślono model zwany regresją wieloraką, bądź wielokrotną, gdzie uwzględniona jest zależność między zmiennymi niezależnymi, a zmienną zależną. Aby wyniki otrzymywane tą metodą były sensowne, należy rozważać takie zbiory danych, gdzie liczba próbek jest przynajmniej pietnaście do dwudziestokrotnie razy większa od liczby zmiennych objaśniających. Zakładając, że rozważamy jaki jest wpływ zbioru n zmiennych na zmienną Y model regresji będzie miał postać:

gdzie:

- parametry modelu (współczynniki regresji) opisujące wpływ i-tej zmiennej - składnik losowy

Niech :

- X oznacza (n \* (k+1)) wymiarową macierz zmiennych objaśnianych zawierających się w n – elementowej próbie na k+1 zmiennych objaśniających

, przy czym (zmienna występująca w wyrazie wolnym):

X=

- y to n – wymiarowy wektor obserwacji pobranych z n– elementowej próby na zmiennych obserwowanych Y:

y=

- oznacza (k+1) wymiarowy wektor kolumnowy współczynników regresji wielorakiej:

- oznacza n-wymiarowy wektor kolumnowy, którego składowymi są tzw. składniki losowe:

zatem mamy:

y= = ∙ +

y = x

Na podstawie macierzy X, oraz wektora y powinniśmy być w stanie obliczyć wektor współczynników regresji, oraz wariancje składnika losowego. Zatem równanie regresji przy n-elementowe próbie i liczbie zmiennych objaśniających równej k ma postać:

więc dla każdej próbki p (p=1...n) mamy:

Podobnie jak w prostej regresji liniowej, parametry szacujemy za pomocą metody najmniejszych kwadratów:

Parametry są to tzw. cząstkowe współczynniki regresji i reprezentują one niezależne wkłady każdej ze zmiennych objaśniających do predykcji porządanej przez nas zmiennej Często w przypadku regresji wielorakiej, bardziej od wartości zmiennej objaśnianej, interesuje nas wpływ poszczególnych zmiennych na wartość końcowego parametru. Niektóre zmienne mogą stymulować, a inne hamować badane zjawisko. Zmienne zaliczające się do grupy z przypadku pierwszego, noszą nazwę stymulantów, te drugie zaś zwane są destymulantami. objaśnianej. Ujemne wartości tych parametrów informują nas o ujemnym oddziaływaniu poziomu zmiennej objaśniającej na objaśnianą, a dodatnie o dodatnim.Istnieją również zmienne neutralne, czyli takie które nie mają żadnego wpływu na badane przez nas zjawisko.

i-ty współczynnik regresji opisuje nam jak średnio zmieniać się będzie wartość zmiennej objaśnianej, gdy i-ta wartość zmiennej X wzrośnie, podczas gdy pozostałe współczynniki będą miały ustaloną wartość. W przypadku kiedy mamy do czynienia z sytuacją, gdzie występuje silna zależność między zmiennymi objaśniającymi, mówimy, że funkcja regresji jest istotna statystycznie. Wtedy okazać się może, że wszystkie zmienne analizowane oddzielnie okazać się mogą nieistotne, powinno się je usunąc z modelu.

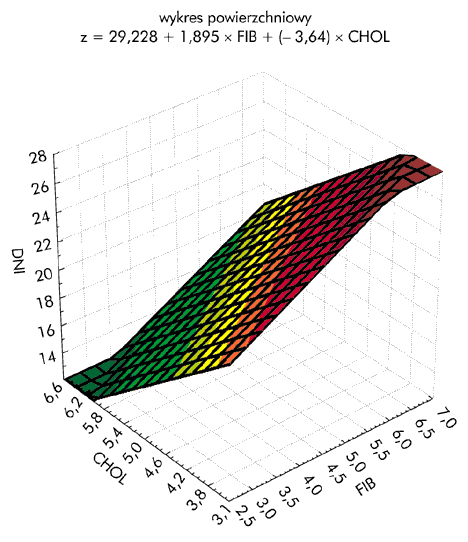
W dobrze skonstruowanym modelu, powinno być jak najwięcej zmiennych objaśniających, które względem siebie są bardzo słabo skorelowane, w przeciwieństwie do poziomu korelacji ze zemienną objaśniającą, który powinen być możliwie jak najwyższy.

Kolejnym elementem jakim powinien charakteryzować się dobrze skonstruowany model regresji wielorakiej jest liniowość związku pomiędzy obserwacjami w próbie na współczynnikach regresji i na zmiennej objaśnianej z dokłądnością do elementu losowego ε.

Spełnony musi być również warunek, mówiący o tym, że żadna zmienna zależna nie może być kombinacją liniową innych zmiennych objaśniających tzw. brak współliniowości.

Założenia jakie musi spełniać składnik losowy także są jasno określone. Mianowicie, musi on mieć wartość oczekiwaną równa zeru, oraz jego wariancja musi być jednakowa dla każdej obserwacji. Składniki losowe muszą być ze sobie nieskorelowane, czyli przy każdej próbie n składniki i są od siebie niezależne, oraz podobnie jak w przypadku regresji liniowej mają rozkład N(0, σ2 ).

Wykresy regresji wielorakiej ze względu na większą ilość parametrów, przyjmują skomplikowane formy:



Rysunek 10 Przykładowy wykres regresji wielorakiej

### 4.2.4 Regresja krokowa

Po zapoznaniu się z modelem regresji wielorakiej, możemy przyjżeć się kolejnemu typowi regresji, jakim jest regresja krokowa. Model ten również opiera się na więcej niż jednej zmiennej objaśniającej, z tym, że początkowo liczba tych parametrów nie jest nam do końca znana. W regresji krokowej z każdym „krokiem” dodajemy, bądź odejmujemy kolejne parametry, dążąc do tego aby jakość naszego modelu była jak najlepsza. Model ten można realizować na dwa soposby:

- Regresja krokowa postępowa, która polega na rozpoczynaniu konstrukcji modelu od takiego, który zawiera jedynie stałą. Następnie kolejno dodawane są predyktory, których wpływ na zmienna objaśniającą jest istotny.

- Regresja krokowa wsteczna, przeciwieństwo metody postępowej. Mianowicie budowe modelu rozpoczynamy od wprowadzenia wszystkich analizowanych zmiennych objaśniających i kolejno usuwamy najmniej istotne z nich.

Przydatność poszczególnych parametrów, możemy ocenic przy pomocy kryterium informacyjnego Akaikiego. Jest to jedna z metod do analizy przydatności modelu regresyjnego o róznej liczbie zmiennych objaśniających. Polega ona na wybraniu takiego modelu o najmniejszej wartości współczynnika:

gdzie:

- to estymowane prawdopodobieństwo, przy założeniach danego modelu, uzyskania takiej właśnie wartości obserwacji j jaka była naprawdę uzyskana.

- q\; to liczba parametrów modelu

W powyższym przypadku nie uwzględniana jest wielkość próby, dlatego do kryterium Akaikiego została dodana poprawka:

gdzie:

- n to liczba prób

Wadą tej metody doboru modelu jest to, że wybiera ona modele o dużej liczbie zmiennych objaśniających, co nie zawsze jest porządane. Dużą zaletą jednak jest wysoki współczynnik precyzji predykcji.

Drugim sposobem na dobór predyktorów jest Bayesowskie kryterium informacyjne Schwartza, które opisane jest wzorem:

Regresja krokowa może być użyteczna w przypadku gdy posiadamy dużą ilość parametrów mających wpływ na badane zjawisko, ale nie chcemy korzystać ze wszystkich, ponieważ ich wprowadzanie absorbowałoby zbyt dużą ilość czasu. Przy omawianiu regresji wielorakiej, była mowa o tym, że aby uzyskać sensowne wyniki, do modelu należy wprowadzić conajmniej 15 razy więcej prób, niż mamy predyktorów, a nie zawsze mamy dostęp do . To może być kolejnym powodem do skorzystania z modelu regresji krokowej, gdyż ograniczy nam ona zmienne objaśniające, tylko do tych, które tak naprawde mają jakiś wpływ na predykowaną wartość.

### 4.2.5 Regresja składowych głównych

Przy stosowaniu regresji składowych głównych, pozbywamy się kłopotu ze skorelowaniem zmiennych objaśniających. W zastępstwie za oryginalne predyktory, podstawiamy składowe główne, które są liniowymi kombinacjami oryginalnych zmiennych, nie będących w korelacji. Od strony praktyczniej, użytych jest tylko kilka początkowych zmiennych, które odzwierciedlają zmienność początkowych predytkorów. W tym przypadku problematyczne może być to, że po usunięciu składowych wytypowanych jako zbędne, nie mamy pewności, że nie spowodowaliśmy przez to jakiegoś zaburzenia, ponieważ pozostawione przez nas zmienne mogą nie być maksymalnie skorelowane z predykowaną zmienną.

### 4.2.6 Regresja częściowych najmniejszych kwadratów

Do rozwiązania problemu opisanego w regresji składowych głównych, pomocna może być regresja częściowych najmniejszych kwadratów. W modelu tym, nasze nowe predyktory wyznaczane są w taki sposób, aby miały jak najwyższy poziom skorelowania ze zmiennymi predykowanymi, przy zachowaniu dobrego wyjaśniania zmienności oryginalnych zmiennych. Metoda ta używana jest przy analizowaniu wplywu dużej ilości zmiennych objaśniających na zmienną objaśnianą, jej dużą zaletą jest użyteczność w przypadku gdy liczba predyktorów jest większa niż liczba przeprowadzonych prób.

### 4.2.8 Regresja nieliniowa

Zdarzyć się może, że pomiędzy predyktorami, a zmiennymi predykowanymi, zachodzić będzie związek nieliniowy. Wtedy z pomocą przychodzi nam model regresji liniowej, który dzieli się na trzy kategorie:

- modele w których nieliniowość występuje względem zmiennych objaśniających, ale nie względem zmiennych objaśnianych:

Taki model można prosto przekształcic do modelu liniowego, dokonujemy podstawienia:

i otrzymujemy:

- Modele nieliniowe w stosunku do predyktorów jak i zmiennych predykowanych, dla których możiwe jest przekształcenie do modelu liniowego z postaci:

Następnie logarytmujemy obie strony:

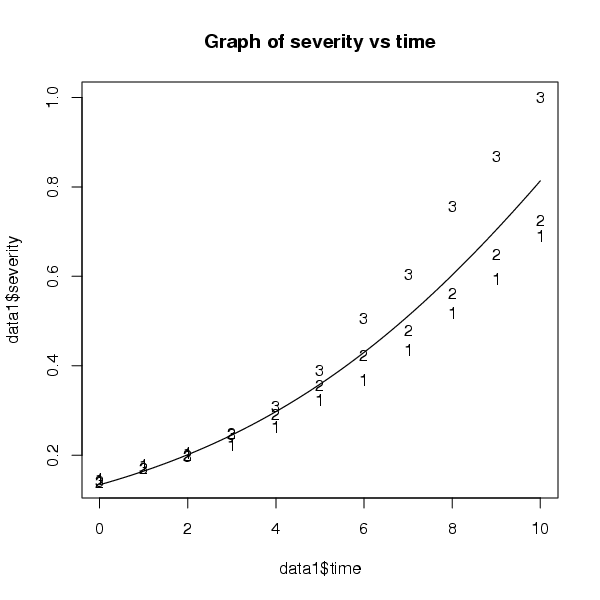
ln +

-Modele całkowicie nieliniowe, czyli takie, gdzie nieliniowość występuje dla predyktorów i zmiennych predykowanych, dla których nie ma możliwości modyfikacji do modelu liniowego:

Nie da się go, w żaden sposób sprowadzić do postaci modelu liniowego, ani poprzez przekształcenie zmiennych objaśnianych jak i zmiennych objaśniających.

Nie jest to jednak wielkim problemem, ponieważ zawsze zaleca się stosowanie modelu nieliniowego niż jego przekształcenie do postaci liniowej i estymacji zmiennych w ten sposób.

Przykładowy wykres regresji nieliniowej prezentuje się następująco:



Rysunek 11 Przykładowy wykres regresji nieliniowej

### 4.2.9 Regresja nieparametryczna

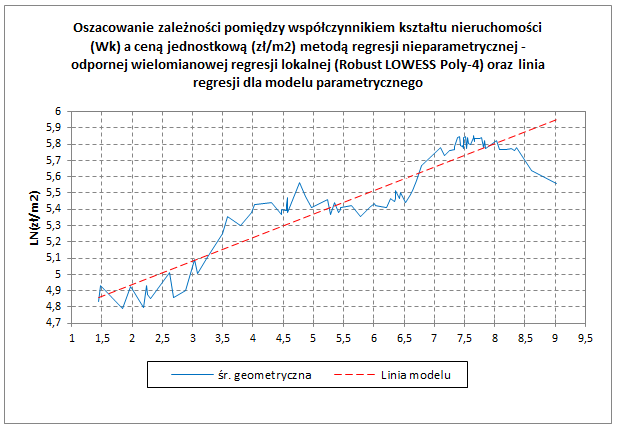
Rozwiązanie to stosuejmy, gdy nie możemy dostosować żadnego innego modelu regresji do przedstawionego nam problemu. Jest on ciekawy i elastyczny ponieważ używać go możemy tam, gdzie nie jesteśmy zainteresowani znajomością parametrów zależności funkcyjnej. W takim przypadku wyznaczamy linię trendu:

- Dzielimy zbiór wartości funkcji na podrozdziały, a na każdym z nich dopasowujemy regresję wieloraką.

- Stosujemy wygładzanie jądrowe

- Regresja najbliższych sąsiadów, gdzie wybierany jest parametr k, wskazujący jaka cząstka danych służyć będzie do zbudowania modelu regresji liniowej. Do wyceny wartości używamy obserwacji:

Przykładowy wykres regresji nieparametrycznej prezentuje się następująco:

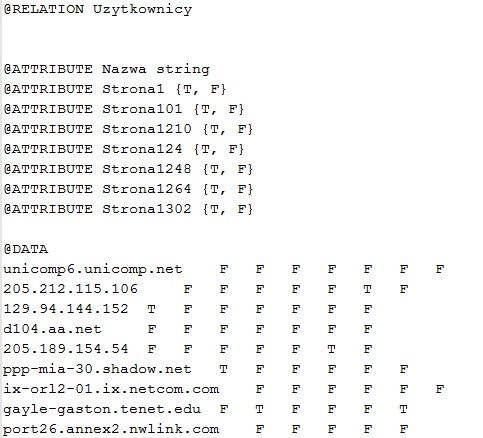


Rysunek 12 Przykładowy wykres regresji nieparametrycznej

## 4.3 Zastosowanie regresji do klasyfikowania danych

W praktyce regresja wykorzystywana jest w wielu dziedzinach takich jak: biologia(przewidywanie prawdopodobieństwa wystąpienia choroby), farmakologia(oszacowanie skuteczności działania leku i wyznaczenie czynników mających wpływ na jego działanie) czy bankowość(ocena wiarygodności kredytobiorcy). To jest tylko kilka przykładów wykorzystywania regresji, w realnym życiu.

Do tworzenia modeli regresji zostało stworzonych sporo narzędzi między innymi oprogramowania takie jak WEKA, czy STATISTICA. Do rozwiązywania problemów z analizą danych stworzony został również specjalny język programowania R. Minusem takiego oprogramowania jest to, że aby moć wprowadzić dane do praogramu w celu ich analizy, należy je odpowiednio przygotować i sformatować do pliku .csv. Poniżej przykładowy poprawnie sformatowany plik z danymi dla oprogramowania WEKA



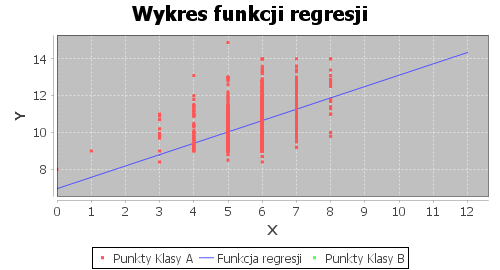
Rysunek 13 Przykładowy plik WEKA

W przypadku języka R, osoba chcąca przeprowadzić jakieś badanie, dodatkowo musi nauczyć się składni, na co poświęci dodatkowy cenny czas.

Modele regresji pozwalają nam wyestymować pożądaną przez nas wartość. Lecz większość użytkowników oczekuje, że otrzymywane wyniki będą również w jakiś spośob sklasyfikowane. Model regresji może sprawdzić się w predykcji jak i klasyfikacji danych. Na potrzeby tej pracy powstała aplikacja, która pozwala nam wyestymować jak i sklasyfikować pewne zbiory danych. Przykładowa klasyfikacja może odbywać się przy pomocy funkcji regresji. W zależności od tego jak bardzo złożony jest model regresji, możemy podzielić dane na dwie lub więcej klas.

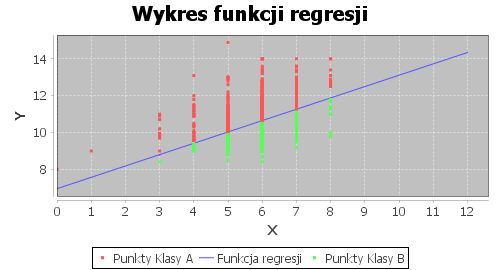
Prostym przykładem klasyfikacji może być równanie funkcji regresji liniowej, gdzie przypisujemy obiekty do klas w zależności od ich położenia względem wykresu funkcji. Jedna klasa to te obiekty, które znajdują się nad lub na linii wyznaczającej funkcję, druga zaś to obiekty leżące pod linią.

Przykład pochodzący z naszej aplikacji przed klasyfikacją:



Rysunek 14 Wykres regresji liniowej przed klasyfikacją

Oraz przykład pochodzący z naszej aplikacji po dokonanej klasyfikacji względem prostej regresji:



Rysunek 15 Wykres regresji liniowej po klasyfikacji

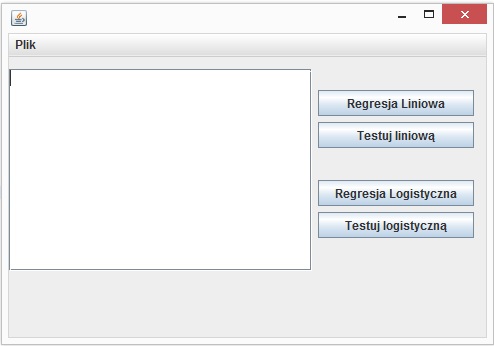
Jak widać na powyższych wykresach, regresja może stanowić podstawę do klasyfikowania danych i radzi sobie z tym conajmniej zadowalająco. Więcej wyników z analizy i klasyfikowania danych znajduje się w rozdziale poświęconym przeprowadzonym badaniom.

# 5. Implementacja

W rozdziale tym, zostanie zaprezentowane aplikacja, którą zaimplementowano w celu przeprowadzenia badań niezbędnych do zrealizowania celu pracy.

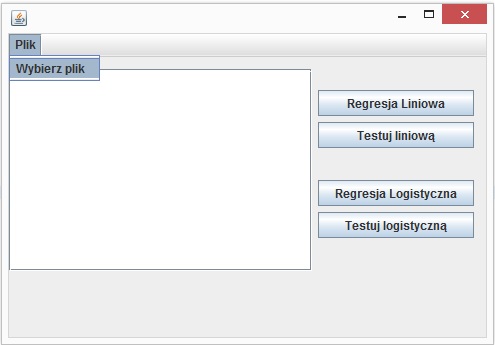
Oprogramowanie zostało stworzone w języku java i spełnia ono podstawowe funkcje potrzebne do analizowania danych przy udziale regresji.

Zaimplementowane zostały dwa algorytmy regresji, jeden dotyczy regresji liniowej, drugi zaś interpretuje wyniki przy pomocy regresji logistycznej. Program charakteryzuje się mało skomplikowanym interfejsem graficznym:



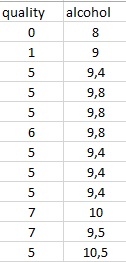
Rysunek 16 Interfejs aplikacji

Dużym plusem naszej aplikacji jest sposób wprowadzania danych uczących lub testowych. Kiedy chcemy wyznaczyć model regresji dla posiadających przez nas danych, wystarczy podać je w postaci pliku .xlsx, który jest bardzo popularnym formatem obsługiwanym przez Excel.



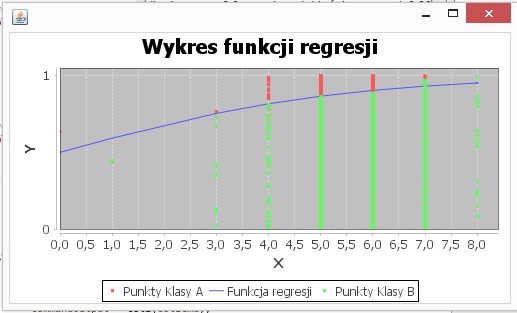
Rysunek 17 Wybieranie pliku z danymi

Przykładowy plik z danymi prezentuje się następująco:



Rysunek 18 Przykładowy plik z danymi

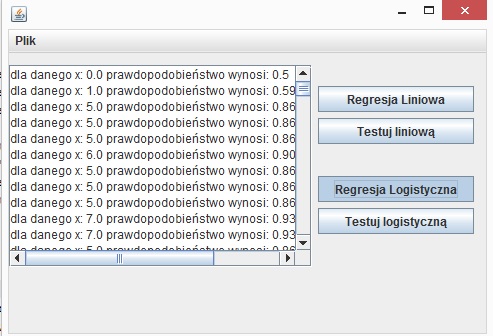
Zawiera on zależność pomiędzy jakością wina a jego mocą. Należy dodać, że w pierwszej kolumnie powinna znajdować się zmienna objaśniająca, a w drugiej zmienna objaśniana, a aplikacja działa tylko dla jednej zmiennej objasniającej. Po wprowadzeniu danych możemy wybrać czy chcemy skorzystać z modelu regresji liniowej, czy regresji logistycznej klikając w odpowiedni przycisk. Wywoła to proces tworzenia modelu regresji, a po jego zakończeniu otrzymamy gotowy wykres regresji, stworzony przy użyciu biblioteki JfreeChart:



Rysunek 19 Wykres regresji logistycznej

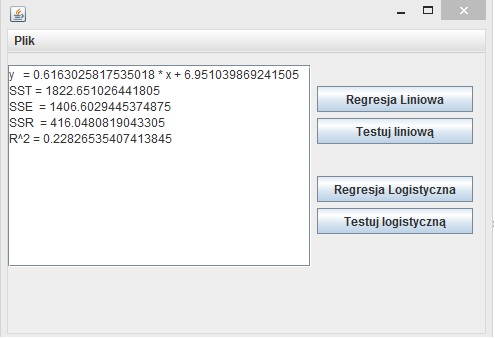
Na którym mamy linie regresji, wyznaczoną przy pomocy wyliczonej funkcji regresji oraz sklasyfikowane punkty, które były podane jako dane uczące.

Dodatkowo w oknie programu mamy wypisane prawdopodobieństwa dla poszczególnych zmiennych objaśniających:



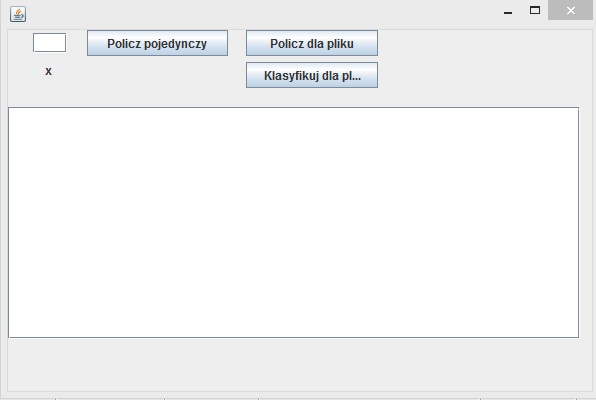
Rysunek 20 Wynik regresji logistycznej

Oraz wzór funkcji, wartości błedów i współczynnika dla regresji liniowej:



Rysunek 21 Wynik regresji liniowej

Po wyliczeniu modelu, możemy przystąpić do testowania go na zbiorze testowym – robimy to w jednej z sekcji testowania (klikamy w przycisk Tesuj iniową lub Testuj logistyczną). Ukaże nam się okno testowania:



Rysunek 22 Okno do testowania wyznaczonego modelu

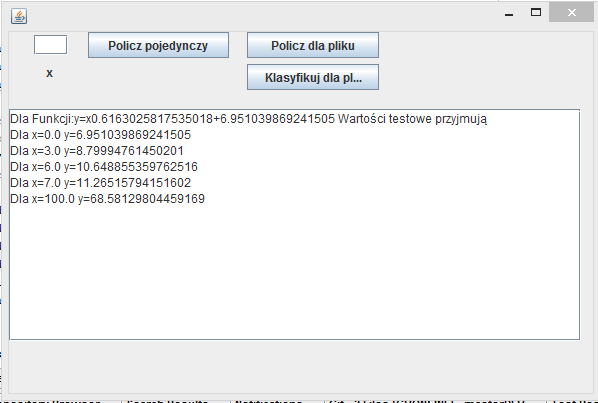
Aplikacja zapamiętuje ostatnio utworzony model, także po ponownym jej uruchomieniu możemy przejść odrazu do testowania naszego zbioru. Testowanie umożliwione jest na dwa sposoby: wpisanie pojedynczej zmiennej, oraz pobranie danych do przetestowania z pliku, który powinien spełniać warunki:

- typ pliku: xlsx

- zawiera jedną kolumne z wartościami x

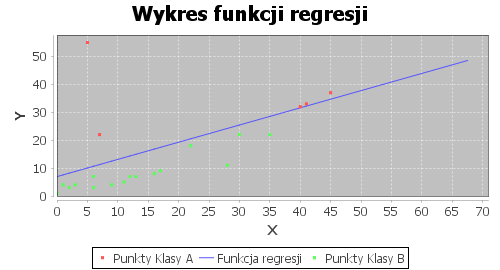
- dane powinny być w formacie liczbowym, jeżeli mają być to ułamki, powinny być oddzielone przecinkiem (tak aby Excel traktował je jako zmienne przecinkowe).

Po zatwierdzeniu wyboru pliku zostaną obliczone przewidzane wartości zmiennej predykowanej, dla zmiennych testowych:



Rysunek 23 Rezultat dla zmiennych testowych

Dane testowe możemy również sklasyfikować, tak jak to miało miejsce dla danych uczących, poniżej przykładowy rezultat klasyfikacji małej ilości danych testowych:

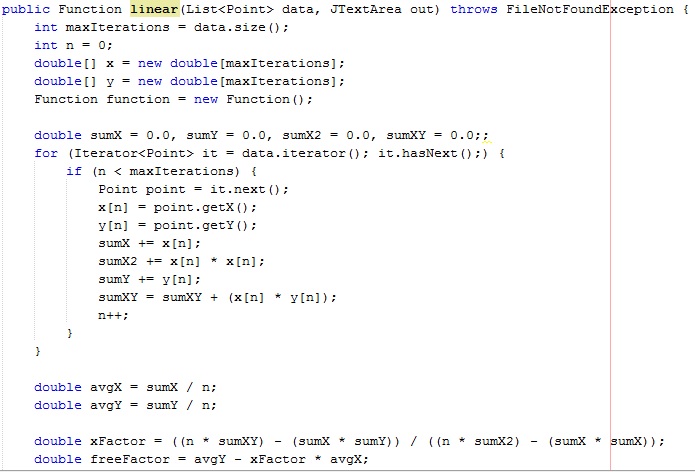


Rysunek 24 Klasyfikacja punktów testowych

Dane do klasyfikacji również podajemy w postaci pliku .xlsx z tą różnicą, że w pierwszej kolumnie znajduje się współrzędna X, w drugiej zaś Y.

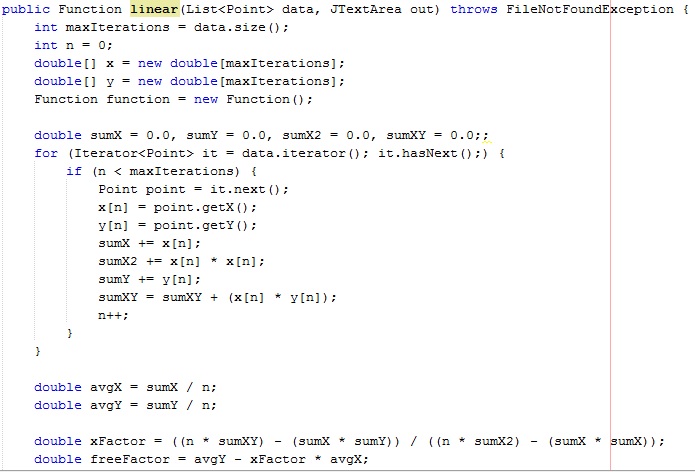
Taka funkcjonalność pozwala nam na przeprowadzenie badań, które przedstawione są w kolejnym rozdziale.

Poniżej przedstawione są dwa fragmenty kodu aplikacji.Pierwszym z nich jest fragment przedstawiający algorytm obliczania współczynników regresji przy użyciu metody najmniejszych kwadratów.



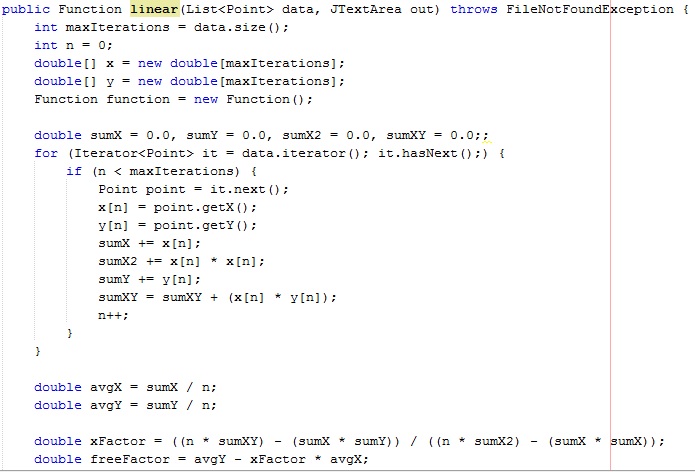
Rysunek 25 Algorytm regresji liniowej

Na początku deklarujemy niezbędne zmienne, następnie obliczamy potrzebne nam sumy wszystkich zmiennych x, y oraz x podniesionego do kwadratu, gdzie n to liczba danych, które wprowadzone zostały z pliku. Tworzone są równierz dwie tablice zawierające podane z pliku zmienne x i y:



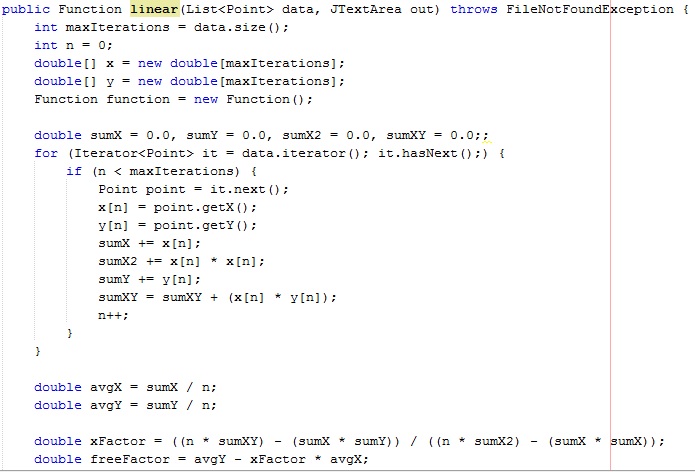
Rysunek 26 Obliczanie sum

Następnie liczymy średnią wartość x oraz y:



Rysunek 27 Obliczanie średnich wartości

a na końcu obliczamy wartości :

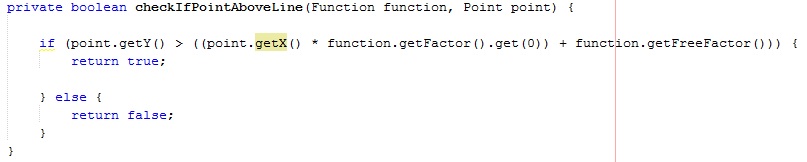


Rysunek 28 Obliczenie wartości końcowych

Odpowiada to wzorom:

=

Kolejnym fragmentem kodu jest algorytm odpowiedzialny za klasyfikowanie punktów:



Klasyfikacja realizowana jest tutaj przy pomocy równania regresji, które zawarte jest w zmiennej function. Zawiera ona współczynniki . Do sprawdzenia czy punkt leży nad prostą podstawiamy jego wartości do równania liniowego a następnie sprawdzamy czy wartość y klasyfikowanego punktu jest większa od wartości x pomnożonej przez współczynnik powiększonej o wolny współczynnik .

# 6. Przeprowadzone badania

## 6.1 Przedstawienie badań

## 6.2 Analiza wyników

# 7. Podsumowanie i wnioski

# 8. Bibliografia

# 9. Spis ilustracji

[Rysunek 1 Diagram woronoja 8](#_Toc443984514)

[Rysunek 2 Działanie klasyfikatora bayesowskiego 9](#_Toc443984515)

[Rysunek 3 Klasyfikator bayesowski biorący pod uwagę liczbę sąsiadów 10](#_Toc443984516)

[Rysunek 4 Neuron 13](#_Toc443984517)

[Rysunek 5 Wielowarstwowa sieć neuronowa 14](#_Toc443984518)

[Rysunek 6 Metoda wektorów nośnych - liniowy podział 15](#_Toc443984519)

[Rysunek 7 Metoda wektorów nośnych - nieliniowy podział 15](#_Toc443984520)

[Rysunek 8 Przykładowy wykres regresji liniowej 18](#_Toc443984521)

[Rysunek 9 Przykładowy wykres regresji logistycznej 20](#_Toc443984522)

[Rysunek 10 Przykładowy wykres regresji wielorakiej 25](#_Toc443984523)

[Rysunek 11 Przykładowy wykres regresji nieliniowej 28](#_Toc443984524)

[Rysunek 12 Przykładowy wykres regresji nieparametrycznej 29](#_Toc443984525)

[Rysunek 13 Przykładowy plik WEKA 30](#_Toc443984526)

[Rysunek 14 Wykres regresji liniowej przed klasyfikacją 31](#_Toc443984527)

[Rysunek 15 Wykres regresji liniowej po klasyfikacji 31](#_Toc443984528)

[Rysunek 16 Interfejs aplikacji 32](#_Toc443984529)

[Rysunek 17 Wybieranie pliku z danymi 33](#_Toc443984530)

[Rysunek 18 Przykładowy plik z danymi 33](#_Toc443984531)

[Rysunek 19 Wykres regresji logistycznej 34](#_Toc443984532)

[Rysunek 20 Wynik regresji logistycznej 34](#_Toc443984533)

[Rysunek 21 Wynik regresji liniowej 35](#_Toc443984534)

[Rysunek 22 Okno do testowania wyznaczonego modelu 35](#_Toc443984535)

[Rysunek 23 Rezultat dla zmiennych testowych 36](#_Toc443984536)

[Rysunek 24 Klasyfikacja punktów testowych 36](#_Toc443984537)

[Rysunek 25 Algorytm regresji liniowej 37](#_Toc443984538)

[Rysunek 26 Obliczanie sum 37](#_Toc443984539)

[Rysunek 27 Obliczanie średnich wartości 38](#_Toc443984540)

[Rysunek 28 Obliczenie wartości końcowych 38](#_Toc443984541)