Uniwersytet im. Adama Mickiewicza w Poznaniu
Wydział Matematyki i Informatyki

Praca magisterska

Klasyfikacja funduszy inwestycyjnych z wykorzystaniem metod nauczenia maszynowego

Classification of Investment Funds by Means of Machine Learning Methods.

Dawid Klimek

nr albumu: 384172
kierunek: informatyka

Promotor:
prof. UAM dr hab. Krzysztof Jassem

Poznań, 2017

Oświadczenie

Ja, niżej podpisany Dawid Klimek student Wydziału Matematyki i Informatyki Uniwersytetu im. Adama Mickiewicza w Poznaniu oświadczam, że przedkładaną pracę dyplomową pt: Klasyfikacja funduszy inwestycyjnych z wykorzystaniem metod nauczenia maszynowego napisałem samodzielnie. Oznacza to, że przy pisaniu pracy, poza niezbędnymi konsultacjami, nie korzystałem z pomocy innych osób, a w szczególności nie zlecałem opracowania rozprawy lub jej części innym osobom, ani nie odpisywałem tej rozprawy lub jej części od innych osób.

Oświadczam również, że egzemplarz pracy dyplomowej w wersji drukowanej jest całkowicie zgodny z egzemplarzem pracy dyplomowej w wersji elektronicznej. Jednocześnie przyjmuję do wiadomości, że przypisanie sobie, w pracy dyplomowej, autorstwa istotnego fragmentu lub innych elementów cudzego utworu lub ustalenia naukowego stanowi podstawę stwierdzenia nieważności postępowania w sprawie nadania tytułu zawodowego.

[ ]* - wyrażam zgodę na udostępnianie mojej pracy w czytelni Archiwum UAM

[ ]* - wyrażam zgodę na udostępnianie mojej pracy w zakresie koniecznym do ochrony mojego prawa do autorstwa lub praw osób trzecich

*Należy wpisać TAK w przypadku wyrażenia zgody na udostępnianie pracy w czytelni Archiwum UAM, NIE w przypadku braku zgody. Niewypełnienie pola oznacza brak zgody na udostępnianie pracy.

.....

(czytelny podpis studenta)

Streszczenie

Celem niniejszej pracy magisterskiej jest przedstawienie zasad wybranych metod uczenia maszynowego oraz stworzenie programu, który wykorzystuje jeden z algorytmów do klasyfikacji funduszy inwestycyjnych.

W pracy zawarto: opis działania naiwnego algorytmu Bayesa oraz sieci neuronowej, zasadę działania i budowę programu stworzonego w ramach projektu autorskiego.

Abstract

The purpose of the thesis is to present selected machine learning methods and a program, which performs the classification of investment funds.

The thesis includes: a short review of the Naive Bayes algorithm and the artificial neural network, a description of a classification tool and its interface as well as the evaluation of the results returned by the application.

Spis treści

Wstęp

Zakres pracy

W pracy została przeprowadzona dyskusja na temat funduszy inwestycyjnych i wybranych metod uczenia maszynowego. Opisany został proces automatycznej klasyfikacji funduszy inwestycyjnych wykorzystujący metody uczenia maszynowego takie jak: naiwny algorytm Bayesa i sieć neuronowa.

Cele pracy

Celem pracy jest wykorzystanie metod uczenia maszynowego w celu klasyfikacji funduszy inwestycyjnych na podstawie danych tekstowych publikowanych przez towarzystwa funduszy inwestycyjnych. Efektem końcowym pracy jest aplikacja automatycznie klasyfikująca fundusze według wprowadzonych przez użytkownika opisów.

Układ pracy

Rozdział 1. przedstawia ogólne informacje o funduszach inwestycyjnych i ich znaczeniu dla człowieka. Rozdział 2. zawiera opisy metod uczenia maszynowego. W rozdziale 3. są zawarte dotychczasowe osiągnięcia i ukazane możliwości wykorzystania uczenia maszynowego w życiu codziennym. W rozdziale 4. opisano zastosowaną metodologię oraz przedmiot aplikacji i jej sposób działania. Rozdział 5. przedstawia ewaluację i podsumowanie pracy.

10:3014353713

Rozdział 1

Fundusze inwestycyjne

Fundusze inwestycyjne to narzędzia inwestowania środków finansowych poprzez zbieranie kapitału od inwestorów i inwestowanie ich w akcje, obligacje, rynek pieniężny lub inne papiery wartościowe. Wszystkie aktywa kupione za pieniądze inwestorów tworzą portfel funduszu, który jest zarządzany przez doradcę finansowego. Inwestorzy w zamian za swój kapitał otrzymują jednostki uczestnictwa reprezentujące prawa majątkowe. Fundusze inwestycyjne są prawnie zobowiązane do wyceny jednostek uczestnictwa każdego dnia roboczego oraz do ich wykupu na życzenie inwestora.

1.1 Rola funduszy inwestycyjnych na rynkach finansowych i w gospodarce

Fundusze inwestycyjne stanowią jeden z podstawowych segmentów rynku finansowego. Dzięki swoim zaletom, dostępności i różnorodności stanowią ważne miejsce wśród form alokowania oszczędności czy też wolnych środków inwestycyjnych. Na rynku finansowym fundusze pełnią funkcje takie jak:

- pośrednik finansowy,
- inwestor instytucjonalny,
- emitent.

Fundusze inwestycyjne jako pośrednik finansowy pośredniczą między podmiotami poszukującymi kapitału inwestycyjnego a podmiotami posiadającymi jego nadmiar z zamiarem alokacji.

Rola inwestora instytucjonalnego w funduszach inwestycyjnych spełniana jest poprzez alokację środków pieniężnych, otrzymanych od uczestników, w instrumenty dostępne na rynku finansowym.

Fundusze inwestycyjne stają się emitentami przez zbywanie tytułów uczestnictwa lub emitowanie certyfikatów inwestycyjnych. Dzięki roli emitenta fundusze zbiorowego inwestowania wzbogacają formę alokacji kapitału przyciągając w ten sposób potencjalnego inwestora.

Kapitał gromadzony, a następnie inwestowany w akcje, obligacje czy papiery dłużne, prowadzi do obniżenia kosztów pozyskania kapitału przez te podmioty.

Dzięki tej operacji pobudzany jest wzrost gospodarczy, a kumulacja oszczędności zapobiega wzrostowi inflacji. Trzeba także zwrócić uwagę na fakt, że fundusze zbiorowego inwestowania cechują się większą rentownością niż lokaty czy depozyty bankowe.

Reasumując, im większe zainteresowanie wśród inwestorów indywidualnych funduszami zbiorowego inwestowania, tym większy poziom zaufania społecznego do rynku finansowego.

Powierzanie oszczędności różnym funduszom inwestycyjnym jest odzwierciedleniem zrozumienia, że w dłuższym horyzoncie czasowym fundusze są bezpiecznym sposobem inwestowania swoich pieniędzy.

1.2 Zalety i wady funduszy inwestycyjnych

Każdy inwestor wybierający fundusze inwestycyjne jako formę inwestycji, pomimo ich zalet powinien zdawać sobie sprawę z ewentualnych zagrożeń oraz wad.

Zalety funduszy inwestycyjnych:

- dywersyfikacja portfela inwestycyjnego,
- profesjonalne zarządzanie,
- bezpieczeństwo inwestycji,
- płynność inwestycji,
- elastyczność wyboru strategii inwestycyjnej.

Największą zaletą funduszy inwestycyjnych jest dywersyfikacja portfela. Jest to proces zwiększania liczby instrumentów w portfelu inwestycyjnym, w celu redukcji ryzyka inwestycyjnego w stosunku do ryzyka indywidualnego związanego z poszczególnymi instrumentami.

O przewadze inwestowania zbiorowego nad indywidualnym świadczy następna zaleta, którą jest profesjonalne zarządzanie. Towarzystwa funduszy inwestycyjnych zarządzają funduszami inwestycyjnymi dostarczając specjalistycznej wiedzy na temat inwestycji kapitałowych. Dzięki temu mogą dokonywać właściwej oceny sytuacji ekonomicznej i umiejętnie przewidzieć trendy rynku finansowego. Pozwalają w ten sposób zaoszczędzić czas, jaki inwestor musiałby poświęcić na podjęcie decyzji inwestycyjnej.

Ważną zaletą jest bezpieczeństwo inwestycji, wynikające z precyzyjnego uregulowania prawnego i ich nadzorowania przez organ administracji państwowej. Bezpieczeństwo jest zagwarantowane również poprzez ograniczenie ryzyka, które wiąże się z bezpośrednim lokowaniem środków na rynku kapitałów pieniężnych przez indywidualnego inwestora.

Kolejną zaletą jest płynność inwestycji, dzięki której tytuły uczestnictwa funduszu inwestycyjnego mogą szybko i łatwo zostać nabyte lub zbyte przez inwestora.

Zaletą elastyczności wyboru strategii inwestycyjnej jest to, że fundusze inwestycje zapewniają szeroką gamę strategii inwestycyjnych o zróżnicowanym poziomie

ryzyka oraz rentowności w celu sprostania indywidualnym preferencjom potencjalnego inwestora.

Wady funduszy inwestycyjnych:

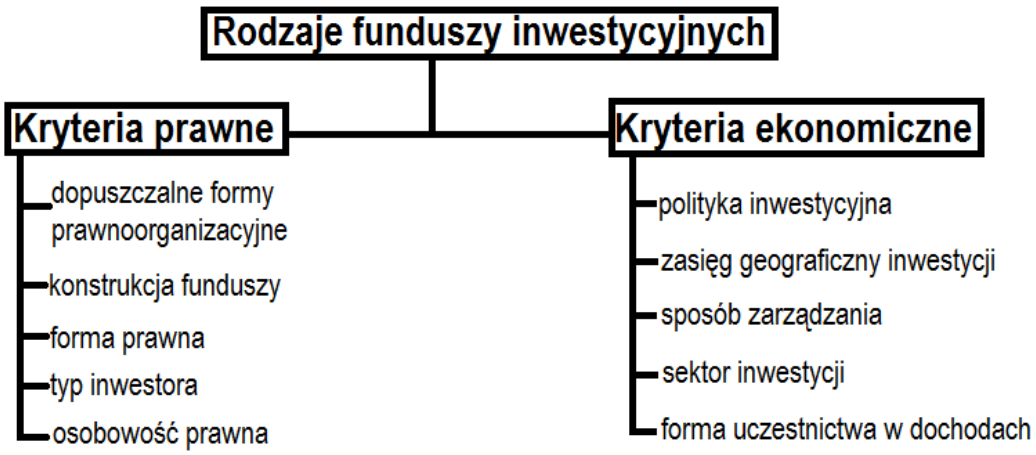
- ryzyko poniesienia straty,
- koszty inwestycji,
- opodatkowanie inwestycji,
- brak bezpośredniego wpływu na politykę inwestycyjną.

Zakup jednostek funduszu inwestycyjnego jest bardziej ryzykowny niż zdeponowanie pieniędzy na rachunku bankowym. Mimo dywersyfikacji portfela oraz profesjonalnego zarządzania, fundusz nie daje żadnej gwarancji na osiągnięcie dochodu.

Koszty inwestycji są niezależne od wyników osiągniętych przez inwestora. Na koszty inwestycji składają się różne opłaty i prowizje, w tym opłata za zarządzanie czy też opłata manipulacyjna. Trzeba pamiętać również o tym, że fundusze inwestycyjne mogą być opodatkowane podatkiem dochodowym lub podatkiem od zysków kapitałowych. Każdy inwestor po ulokowaniu środków w funduszu inwestycyjnym traci możliwość wpływu na strategie inwestycyjne do czasu zmiany funduszu.

1.3 Rodzaje funduszy inwestycyjnych

W zależności od kryterium przyjętego za podstawę podziału fundusze inwestycyjne można klasyfikować na wiele sposobów. Wyróżnia się kryteria prawne i kryteria ekonomiczne (Rysunek 1.1).



Rysunek 1.1: Rodzaje funduszy inwestycyjnych

Kryteria prawne pozwalają scharakteryzować fundusze pod względem konsekwencji prawnych, z jakimi może się spotkać inwestor. Kryteria ekonomiczne wskazują natomiast kształt uprawnień i obowiązków uczestników funduszy w zależności od ekonomicznych zasad ich polityki inwestycyjnej. Polityka inwestycyjna jest uregulowana w statutach towarzystw funduszy inwestycyjnych i prospektach funduszy.

Kryterium dopuszczalnych form prawno-organizacyjnych wynika z regulacji prawnych danego kraju i jest najistotniejszym kryterium prawnym, ponieważ decyduje o innych ujęciach klasyfikacyjnych. Wśród tego kryterium, można wyróżnić fundusze inwestycyjne regulowane - oparte na przepisach o funduszach inwestycyjnych i fundusze inwestycyjne nieregulowane - funkcjonujące na podstawie innych obowiązujących przepisów.

Kolejnym kryterium prawnym funduszy inwestycyjnych jest kryterium konstrukcji funduszu. Na podstawie tego kryterium możemy wyróżnić trzy różne fundusze, t.j.: otwarte, zamknięte i mieszane. Najczęściej występującym rodzajem są fundusze otwarte, które charakteryzują się otwartością swojej działalności, przez co podlegają ciągłym zmianom liczby osób i wartości aktywów. Fundusze zamknięte w porównaniu do otwartych występują jedynie przy zamkniętych kręgach inwestorów. Są one bardziej efektywne od funduszy otwartych, ponieważ mają niższe koszty działalności oraz stały kapitał początkowy. Fundusze mieszane posiadają cechy zarówno funduszy otwartych jak i zamkniętych, umożliwiając prowadzenie bardziej aktywnej polityki inwestycyjnej.

Na rynku inwestycyjnym wyróżniamy również fundusze statutowe i umownie sklasyfikowane według formy prawnej funduszu. Cechą szczególną funduszy statutowych jest brak rozdziału między majątkiem spółki, a majątkiem funduszu, którym spółka zarządza. Odwrotna sytuacja występuje w funduszach umownych, gdzie majątek funduszu nie jest wspólny z majątkiem spółki, lecz ma charakter wydzielonej masy majątkowej i służy jedynie do nabywania innych instrumentów finansowych.

Bardzo ważny w odróżnianiu funduszy inwestycyjnych jest również typ inwestora. Według tego kryterium można wyróżnić trzy typy funduszy: publiczne, specjalne i prywatne. Fundusze publiczne nie są przeznaczone dla konkretnej grupy, a ich uczestnikami mogą być zarówno osoby fizyczne jak i firmy. Fundusze specjalne mają już ograniczone grono uczestników - zazwyczaj są to inwestorzy instytucjonalni dysponujący dużym kapitałem. Natomiast fundusze prywatne są zamknięte, a swoją ofertę kierują jedynie do wybranych inwestorów.

Ostatnim kryterium prawnym funduszy inwestycyjnych jest kryterium osobowości prawnej. Fundusze mogą posiadać osobowość prawną lub jej nie posiadać. Funduszami posiadającymi osobowość prawną są wszystkie fundusze statutowe oraz ograniczona część umownych. Posiadanie osobowości prawnej przez fundusz inwestycyjny zwiększa jego formalność (obowiązek rejestracji), ale za to daje użytkownikom większy wpływ na jego politykę. Fundusz bez osobowości prawnej musi być zrównoważony, dlatego przede wszystkim funduszami bez osobowości prawnej są fundusze umowne.

Kryterium polityki inwestycyjnej, metod alokacji aktywów funduszy i stopnia ryzyka inwestycyjnego dzieli nam fundusze na tradycyjne i alternatywne. Fundusze tradycyjne są dostępne dla drobnych inwestorów przez co podlegają nadzorowi nad rynkiem finansowym. Są to fundusze: akcyjne, hybrydowe, obligacyjne i rynku

pieniężnego. Fundusze alternatywne posiadają niską przejrzystość funkcjonowania i zazwyczaj nie podlegają żadnym regulacjom prawnym.

Ze względu na zasięg geograficzny istnieje podział na fundusze o zasięgu krajowym, inwestujące za granicą oraz bez określonego zasięgu geograficznego. Te ostatnie charakteryzują się największą swobodą działania zarówno w kraju macierzystym jak i poza jego granicami.

Kolejnym kryterium wyróżniającym fundusze inwestycyjne jest kryterium sposobu zarządzania. Fundusze mogą być zarządzane w sposób aktywny. Są one konstruowane na podstawie aktywnej strategii, wśród których najpopularniejszą jest *market timing*[1]. Zarządzanie funduszami może odbywać się również w sposób pasywny, który wiąże się z ponoszeniem znacznie niższych kosztów. Sposób pasywny zarządzania opiera się głównie na zakupie instrumentów finansowych i czekania do momentu ich upłynnienia.

Fundusze inwestycyjne można również podzielić według kryterium formy uczestnictwa inwestorów w dochodach. Wyróżniamy fundusze tezauryzujące (akumulacyjne) oraz fundusze dystrybucyjne (wypłacające). Dochód z funduszy tezauryzujących jest reinwestowany, co daje możliwość zwiększenia całkowitego dochodu, natomiast dochód z funduszy dystrybucyjnych jest wypłacany inwestorom w całości lub częściowo z odsetkami.

Ze względu na kryteria polityki inwestycyjnej wyróżnia się fundusze tradycyjne i fundusze alternatywne.

Wśród funduszy tradycyjnych wyróżnia się:

- fundusze akcyjne,
- fundusze mieszane,
- fundusze dłużne,
- fundusze rynku pieniężnego.

W funduszach akcyjnych większość części portfela inwestycyjnego stanowią akcje. Fundusze te cechują się najwyższym ryzykiem inwestycyjnym spośród wszystkich rodzajów funduszy.

W zależności od poziomu ryzyka inwestycyjnego i celów inwestycyjnych wśród funduszy akcyjnych wyróżnia się:

- fundusze wzrostu,
- fundusze dochodu z kapitału,
- fundusze wzrostu i dochodu z kapitału,
- fundusze indeksowe.

[1] http://finansopedia.forsal.pl/encyklopedia/gielda/hasla/912856,strategia-market-timing.html

Fundusze mieszane inwestują w akcje i obligacje, a proporcje udziału tych papierów wartościowych w portfelu zależą od przyjętej polityki funduszu. Akcje są czynnikiem wpływającym na wzrost wartości zarządzanego portfela, z kolei papiery dłużne stanowią element stabilizujący.

Wśród funduszy mieszanych wyróżnia się:

- fundusze stabilnego wzrostu,
- fundusze zrównoważone.

W funduszach dłużnych w skład portfela wchodzą wyłącznie papiery wartościowe o stałym dochodzie, przede wszystkim obligacje. Cały kapitał jest inwestowany w jeden określony rodzaj obligacji lub dywersyfikowany w kilka rodzajów. Fundusze te cechują się umiarkowanym przyrostem ulokowanego w nich kapitału, przy mniejszym ryzyku strat.

Fundusze rynku pieniężnego inwestują kapitały swoich uczestników w instrumenty dłużne o krótkich terminach zapadalności. Fundusze te cechują się wysoką płynnością oraz niskim lub nawet zerowym ryzykiem inwestycyjnym, dając stabilny dochód o umiarkowanej wysokości.

Fundusze alternatywne różnią się od funduszy tradycyjnych tym, że mogą inwestować zarówno w aktywa finansowe, jak i niefinansowe, przez to cechują się znacznie większym poziomem ryzyka inwestycyjnego niż fundusze tradycyjne. Skład portfela głównie zależy od wyspecjalizowanej polityki inwestycyjnej. Fundusze te są alternatywą dla tradycyjnych funduszy inwestycyjnych, dzięki niskiej korelacji z instrumentami tradycyjnymi.

Wśród funduszy alternatywnych wyróżnia się:

- fundusze hedgingowe (fundusze te nie podlegają ścisłym regulacjom prawnym),
- fundusze private equity/venture capital (fundusze te oferują wsparcie kapitałowe małym, ale prężnie prosperującym firmom),
- fundusze nieruchomości,
- fundusze surowcowe i towarowe,
- fundusze związane z inwestycjami kolekcjonerskimi.

Ze względu na kryterium zasięgu geograficznego wyróżnia się:

- fundusze inwestycyjne o zasięgu krajowym - inwestują kapitał w papiery wartościowe i inne instrumenty na rynku danego kraju,
- fundusze inwestycyjne regionalne - inwestują kapitał w danych region np. Azja, europejskie rynki wschodzące,
- fundusze bez określonego zasięgu geograficznego.

Rozdział 2

Uczenie maszynowe

Uczenie maszynowe[1] (ang. Machine Learning) to dziedzina nauki zajmująca się analizą procesów uczenia się oraz tworzeniem systemów, które doskonalą swoje działanie na podstawie doświadczeń z przeszłości. Jest to dziedzina interdyscyplinarna ze szczególnym uwzględnieniem takich jak: sztuczna inteligencja, informatyka i statystyka. Definicja uczenia maszynowego według Toma Mitchella [5]:

> "Mówimy, że maszyna uczy się zadania T w oparciu o doświadczenie E i miarę jakości P, jeśli wraz z przyrostem doświadczenia E poprawia się jakość wykonywanego zadania T mierzona przez miarę P"

2.1 Naiwny algorytm Bayesa

Naiwny algorytm Bayesa, tak jak sama nazwa wskazuje, jest algorytmem klasyfikującym opartym na teorii Bayesa i stosuje naiwne założenia dotyczące dystrybucji danych. Naiwny algorytm Bayesa używa zbioru treningowego: $D = (x_1; y_1), \cdots, (x_n; y_n)$ gdzie $x_i = (x_{i,1}, \cdots, x_{i,d}) \in R_d$ jest wektorem cech d wymiarowym, a $y_i \in Y = \{1, \cdots, m\}$, gdzie m jest liczbą etykiet klas. Model zakłada istnienie rozkładu $P(x, y)$. Co więcej, naiwnie zakłada się, że każda cecha jest warunkowo niezależna, biorąc pod uwagę klasę y[2]:

$$P(x|y) = \prod_{i=1}^{d} P(x_i|y)$$

Na przykład, człowiek może być uznany za mężczyznę, jeśli jest łysy i stary. Naiwny algorytm Bayesa uważa, że każda z tych cech przyczynia się niezależnie do prawdopodobieństwa, że dana osoba jest mężczyzną, niezależnie od ewentualnych korelacji pomiędzy łysieniem, a wiekiem osoby. Mimo, że to naiwne założenie nie jest zazwyczaj odzwierciedlane w praktyce to upraszcza oszacowanie i umożliwia uzyskanie rozkładu a posteriori[2]. Każda marginalna gęstość $P(x_i|y)$ może być oszacowana przy użyciu dowolnej dystrybucji np. rozkładu Gaussa, wielowymiarowego rozkładu dla dyskretnych funkcji.

Algorytm wylicza kolejne prawdopodobieństwo a posteriori przy użyciu twierdzenia Bayesa:

[1] https://pl.wikipedia.org/wiki/Uczenie_maszynowe

[2] rozkład obliczany na podstawie wyników doświadczenia, czyli częstości.

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Najwyższe prawdopodobieństwo przynależności do y może być wyznaczone za pomocą wzoru:

$$y = argmax_{y \in Y} P(y|x) = argmax_{y \in Y} \frac{P(x|y)P(y)}{P(x)}$$

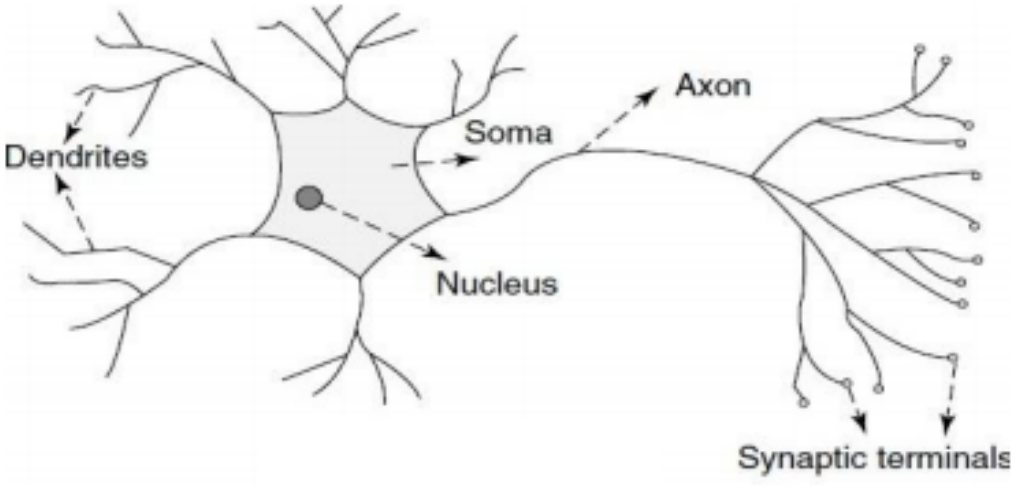
gdzie $P(x)$ jest stałe, co daje nam:

$$y = argmax_{y \in Y} P(x|y)P(y) = argmax_{y \in Y} P(y) \prod_{i=1}^{d} P(x_i|y)$$

Naiwny klasyfikator Bayesa jest oparty na optymistycznych założeniach, które przeważnie są błędne, gdyż w rzeczywistości większość atrybutów opisujących świat jest ze sobą skorelowanych. Pomimo tego, często przewyższa jakością znacznie bardziej złożone alternatywy. Dzięki swojej prostocie jest bardzo szybki w obliczeniach.

2.2 Sieci neuronowe

Sieć neuronowa (sztuczna sieć neuronowa) wywodzi się od biologicznej koncepcji neuronów. Neuron jest podstawową komórką strukturalną w mózgu. Aby zrozumieć sieć neuronową, należy zrozumieć, jak działa neuron. Neuron składa się głównie z czterech części: dendrytów, jądra, somy i aksonów (Rysunek 2.1).

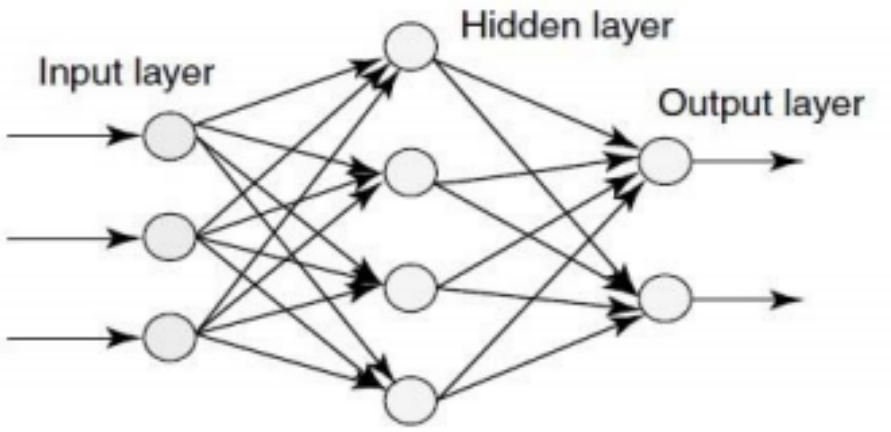


Rysunek 2.1: Neuron człowieka [3]

Dendryty otrzymują sygnały elektryczne, które są analizowane przez som. Wynik procesu jest przenoszony przez akson do terminali dendrytowych, gdzie dane wyjściowe przekazywane są do następnego neuronu. Jądro jest sercem neuronu, a połączone wzajemnie neurony tworzą sieć neuronową, dzięki której impulsy elektryczne są przekazywane po mózgu.

Sztuczna sieć neuronowa zachowuje się podobnie, z tym że składa się z trzech warstw:

- warstwa wejściowa przyjmująca dane wejściowe (podobnie jak dendryty),
- warstwa ukryta przetwarzająca dane wejściowe (soma i akson),
- warstwa wyjściowa wysyłająca obliczony wynik.



Rysunek 2.2: Warstwy sztucznej sieci neuronowej [3]

Do każdego neuronu j jest przypisana funkcja aktywacji l_j. Każda krawędź z węzła j' do j posiada wagę $w_{j'j}$ Wartość v_j każdego neuronu j oblicza się poprzez zastosowanie jego funkcji aktywacji dla sumy iloczynów wag krawędzi łączącej węzły wejściowe:

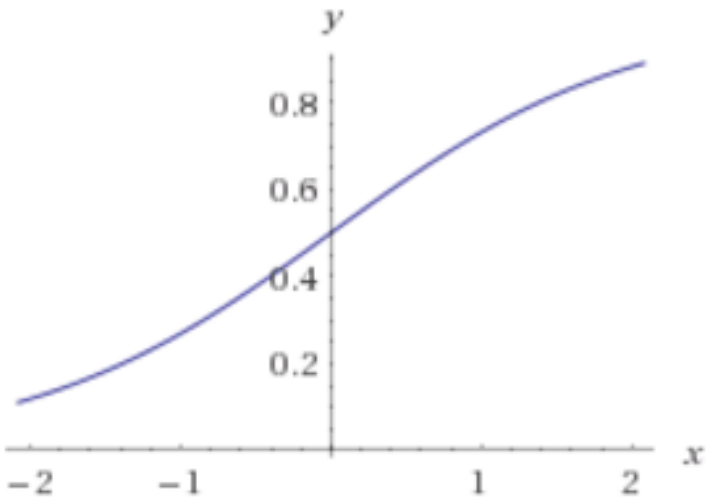
$$v_j = l_j(\sum_{j'} w_{j'j} \cdot v_{j'}).$$

Pożądane cechy funkcji aktywacji:

- ciągłe przejście pomiędzy swoją wartością maksymalną a minimalną,
- łatwość obliczenia oraz istnienie ciągłej pochodnej,
- możliwość wprowadzenia do argumentu parametru α do ustalania postaci analitycznej funkcji aktywacji.

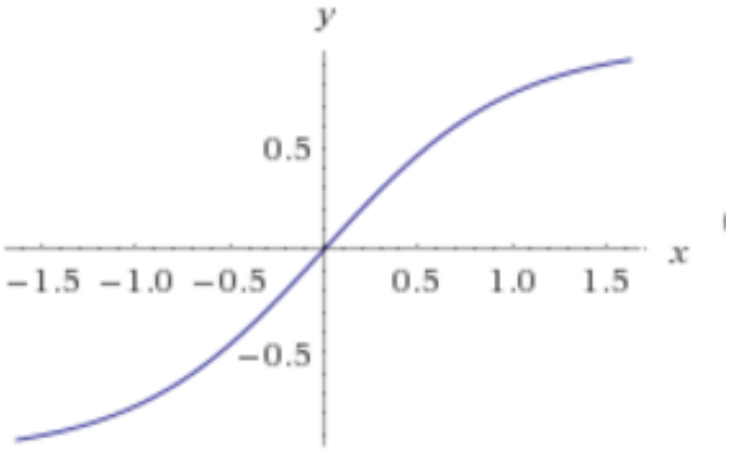
Przykładowe funkcje aktywacji podane są na rysunkach:

- Funkcja aktywacji sigmoid $l(z) = \frac{1}{1+e^{-x}}$



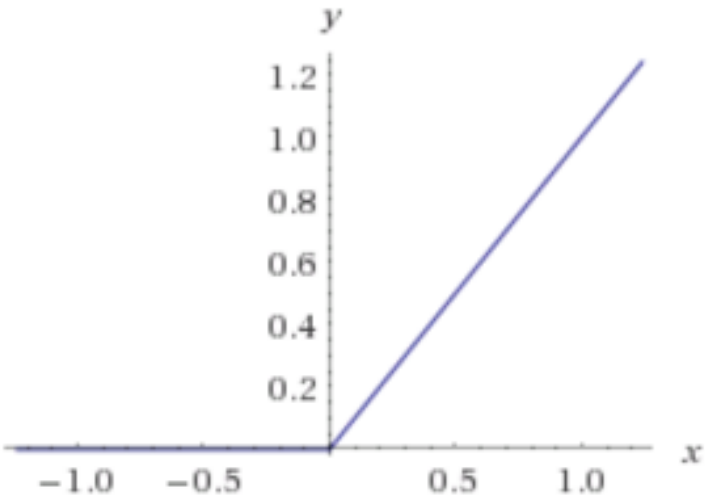
Rysunek 2.3: Funkcja aktywacji sigmoid

- Funkcja aktywacji tanh $l(z) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$



Rysunek 2.4: Funkcja aktywacji tanh

- Funkcja aktywacji ReLU $l(z) = max(0, x)$



Rysunek 2.5: Funkcja aktywacji ReLU

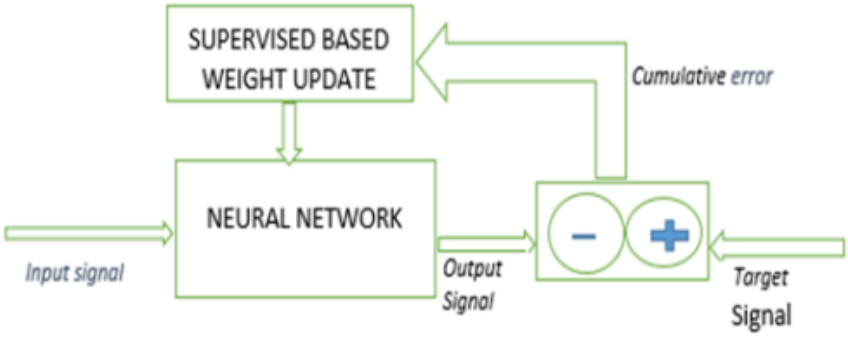
W warstwie wyjściowej sieci neuronowej stosowana jest funkcja softmax:

$$softmax = \frac{e^{v_j}}{\sum e^{v'_j}}$$

Występują głównie trzy rodzaje sztucznej sieci neuronowej [3]:

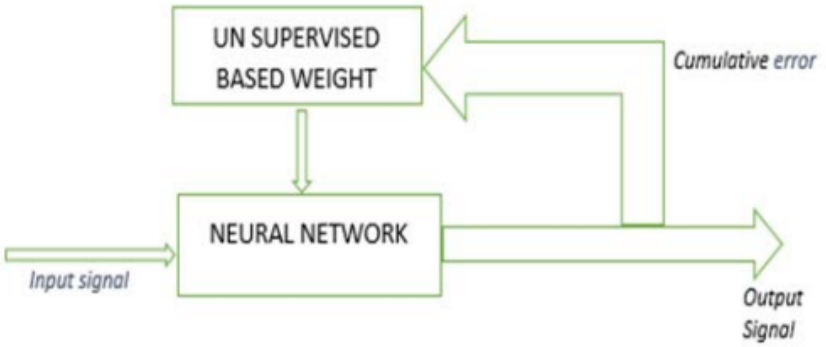
- nadzorowana,
- bez nadzoru,
- wzmocniona.

W nadzorowanej sieci neuronowej dane wejściowe są już znane i przewidywany wynik jest porównywany z rzeczywistym wynikiem. W oparciu o błąd, wagi sieci neuronowej zostają zmienione i ponownie przechodzą przez sieć neuronową. Proces się powtarza aż do osiągnięcia przyjętego maksymalnego błędu (Rysunek 2.6).



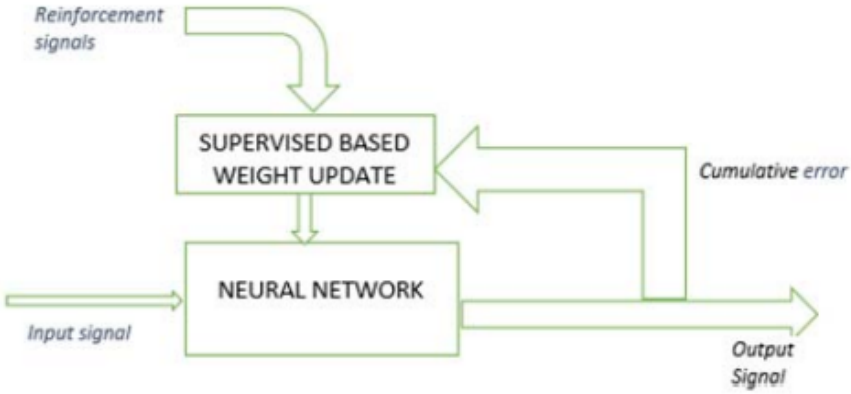
Rysunek 2.6: Schemat nauki sieci neuronowej z nadzorem [3]

W sieci neuronowej bez nadzoru dane wejściowe nie są sklasyfikowane. Sieć stara się kategoryzować dane według podobieństw. Sprawdza ona korelacje między danymi wejściowymi, a utworzonymi grupami w celu uzyskania najmniejszego błędu (Rysunek 2.7).



Rysunek 2.7: Schemat nauki sieci neuronowej bez nadzoru [3]

Wzmocniona sieć neuronowa zachowuje się jak człowiek komunikujący się z otoczeniem. Sieć neuronowa pobiera informacje z zewnątrz, sprawdza czy wynik jest poprawny - jeżeli tak, to wagi sieci neuronowej są wzmacniane (powiększane lub zmniejszane), a w przeciwnym wypadku osłabiane (Rysunek 2.8).

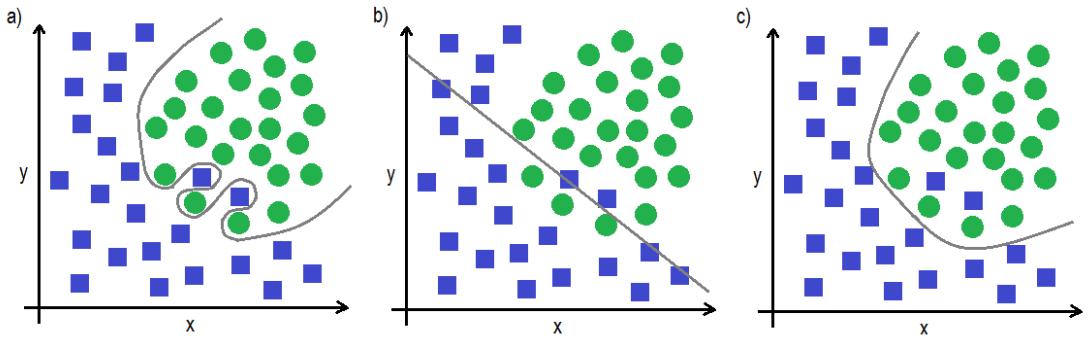


Rysunek 2.8: Schemat nauki wzmocnionej sieci neuronowej [3]

2.3 Generalizacja a przetrenowanie

Celem algorytmu uczenia się maszyn nadzorowanych jest znalezienie modelu, który najlepiej odzwierciedla rozkład danych. Przy konstruowaniu modelu, początkowym celem jest znalezienie hipotezy, która jest najbardziej trafna dla zadanego problemu. Jednak dane wykorzystywane do trenowania modelu rzadko są całkowicie reprezentatywne dla prawdziwego rozkładu. Tworzenie modelu z nadmierną ilością cech, może prowadzić do nadmiernego dopasowania. Zjawisko to nazywane jest przetrenowaniem i ma decydujące znaczenie przy konstruowaniu

modelu, którego celem jest przewidywanie klas dla nowych próbek. Trzeba zatem ograniczyć złożoność modelu, nie czyniąc go zbyt uproszczonym, gdyż prowadzi to do zbyt dużej generalizacji. Przetrenowanie i nadmierna generalizacja to dwie największe przyczyny słabych osiągów algorytmów uczenia maszynowego. Na rysunku 2.9 widzimy przykład: nadmiernej generalizacji, przetrenowania oraz zoptymalizowany model.



Rysunek 2.9: a) przetrenowanie b) nadmierna generalizacja c) zoptymalizowany model

Istnieje kilka sposobów rozwiązania problemu przetrenowania:

- ograniczyć ilość cech - wybierać tylko tych najważniejszych,
- zwiększyć liczbę danych treningowych,
- utworzyć kryterium stopu uczenia się.

Najbardziej oczywistym sposobem na rozwiązywanie problemu jest ograniczenie liczby cech. Cechy organiczna się poprzez wybranie jednej z kilku lub usunięcie tych nieznaczących. Nie zawsze możliwe jest takie rozwiązanie, gdyż dane często są nieliniowe. Każdy typ algorytmu uczenia maszynowego ma również swój własny sposób na rozwiązanie problemu generalizacji i przetrenowania.

Rozdział 3

Zastosowanie metod uczenia maszynowego

W ostatnim czasie badania nad *text mining* czyli eksploracją danych tekstowych stają się coraz ważniejsze z powodu narastającej liczby elektronicznych dokumentów z różnych źródeł. Wyróżnia się dane nieuporządkowane lub częściowo ustrukturalizowane, takie jak: elektroniczne rządowe repozytoria, artykuły informacyjne, biologiczne bazy danych, biblioteki cyfrowe, fora internetowe. Dane wymagają prawidłowej klasyfikacji.

Przetwarzanie języka naturalnego, data mining (eksploracja danych) i uczenie maszynowe jest wspólnie wykorzystywane w celu automatycznej klasyfikacji oraz wyszukiwania wzorców w dokumentów tekstowych. Głównym celem dziedziny *text mining* jest umożliwienie użytkownikowi wyodrębnienia informacji z tekstu wraz z ich klasyfikacją i opisem. Prawidłowe opisanie, przedstawienie i sklasyfikowanie dokumentów wiąże się z problemami takimi jak: odpowiednie zindeksowanie, dobór poprawnej klasyfikacji czy też odporność na przetrenowanie.

Około 90% danych na świecie jest przechowywane w nieustrukturalizowanym formacie[4], dlatego zwiększa się zapotrzebowanie na automatyczne pobieranie dużych volumenów danych i ich analizy w celu wspomagania pracy człowieka. Badanie trendów rynkowych opartych na treściach internetowych artykułów informacyjnych, nastrojach i ważnych wydarzeniach jest głównym zadaniem *text mining*.

W 2006 roku Guobin Ou i Yi Lu Murphey w swojej pracy ukazali wyniki dokładności klasyfikacji danych "Glass" i "Shuttle", wykorzystując różne konfiguracje sieci neuronowych. Omówili sposób nauczania sieci, złożoność, czas uczenia. Stworzyli 6 rożnych architektur sieci neuronowych i zbadali skuteczność na różnych danych. W zależności od danych i użytej architektury sieci neuronowej skuteczność większości przypadków wynosiła ponad 90% [9].

W 2007 roku Qiong Wang, George M. Garrity, James M. Tiedje i James R. Cole użyli naiwnego algorytmu Bayesa do stworzenia klasyfikatora rozpoznającego bakteryjne sekwencje 16S rRNA. Ten klasyfikator został wytrenowany na korpusie posiadającym 28109 sekwencji i klasyfikował nowe próbki z 98% dokładnością [7].

W 2010 roku Pablo D. Robles-Granda i Ivan V. Belik[5] zastosowali klika technik uczenia maszynowego w celu klasyfikacji danych (firm europejskich i japońskich) na podstawie 59 cech finansowych. Dla danych europejskich firm za pomocą drzewa decyzyjnego otrzymali skuteczność klasyfikacji $50,29\%$, dla naiwnego Bayesa - $21,72\%$, a sieci neuronowych - $14,16\%$. Dla danych japońskich firm

za pomocą drzewa decyzyjnego otrzymali skuteczność klasyfikacji - $59,05\%$, dla naiwnego Bayesa $15,53\%$, a sieci neuronowych - $49,23\%$ [5].

W 2011 roku S. L. Ting, W. H. Ip, Albert H. C. Tsang, sprawdzili czy, naiwny algorytm Bayesa sprawdza się w klasyfikacji dokumentów tekstowych oraz opisali proces tworzenia klasyfikatora. Ewaluacja polegała na sprawdzeniu skuteczności algorytmu w różnych wariantach. W najlepszym przypadku klasyfikator oparty na naiwny algorytmie Bayesa uzyskał skuteczność 97%[8] dla klasyfikacji 4000 dokumentów tekstowych.

Rozdział 4

Opis eksperymentu

Głównym celem eksperymentu jest stworzenie programu, który klasyfikuje opis funduszu inwestycyjnego. W wyniku tej klasyfikacji program zwraca: kategorię funduszu, region inwestowania funduszu oraz listę funduszy inwestycyjnych, z bazy danych, o zbliżonej polityce inwestycyjnej. Dzięki tej klasyfikacji program umożliwia zapoznanie się z funduszami dostępnymi na rynku, odpowiadającymi wprowadzonemu opisowi. Osiągnięcie głównego celu projektu jest możliwe poprzez skuteczną klasyfikację wprowadzonego opisu funduszu. W tym celu został przygotowany zbiór danych treningowych i stworzone klasyfikatory oparte na naiwnym algorytmie Bayesa i sieciach neuronowych. Ewaluacja klasyfikatorów wyłoniła najlepszy z nich: klasyfikator o najwyższej skuteczności, który następnie został wykorzystany w projekcie autorskim.

4.1 Opis użytych technologii

Projekt programistyczny bazuje na pewnym zestawie gotowych narządzi takich jak biblioteki, moduły, standardy protokołów itp. W tym rozdziale zostaną opisane podstawowe narzędzia informatyczne użyte podczas eksperymentu.

4.1.1 Jupyter Notebook

Jupyter notebook[1] jest elastycznym narzędziem, które pomaga tworzyć czytelne analizy z wykorzystaniem wykonalnego kodu, obrazów, komentarzy, wzorów, wykresów i innych danych multimedialnych. Dostępny jest jako otwarto-źródłowa aplikacja klient-serwer umożliwiająca edytowanie i uruchamianie dokumentów tekstowych, notatników, za pośrednictwem przeglądarki internetowej. Aplikacja może być wykonywana na komputerze bez dostępu do Internetu lub może być zainstalowana na zdalnym serwerze, w którym można uzyskać dostęp do notatnika przez Internet. Pola tekstowe udostępniane przez Jupyter Notebook to seria komórek zawierających kod wykonywalny lub tekst w formacie markdown[2], popularnym językiem znaczników HTML dla opisów. System również wspiera składnię LaTeX[3] dla równań matematycznych, wykorzystując bibliotekę MathJax[4]. Notatniki

[1] http://jupyter.org/
[2] https://pl.wikipedia.org/wiki/Markdown
[3] https://www.latex-project.org/
[4] https://www.mathjax.org/

można zapisywać i łatwo udostępniać w formacie *.ipynb*, który w gruncie rzeczy jest obiektem JSON[5].

4.1.2 Python

Python[6] jest interpretowanym, interaktywnym językiem programowania stworzonym przez Guido van Rossuma w 1990 roku. Jest rozwijany jako projekt Open Source, zarządzany przez Python Software Fundation. Python posiada w pełni dynamiczny system typów i automatyczne zarządzanie pamięcią. Python jest stosunkowo łatwy do nauki, ponieważ wymaga unikalnej składni, która koncentruje się na czytelności. Dodatkowo, Python obsługuje wykorzystanie obiektów i bibliotek, co oznacza, że programy mogą być zaprojektowane w stylu modułowym i kod może być ponownie wykorzystany w różnych przedsięwzięciach. Opracowane moduły, można skalować do innych projektów i z łatwością je importować lub eksportować. Największą zaletą języka Python jest to, że biblioteka standardowa i interpreter są dostępne bezpłatnie w formie binarnej i źródłowej. Python i wszystkie niezbędne narzędzia są dostępne na wszystkich głównych platformach. Dlatego jest to atrakcyjna opcja dla programistów, którzy nie chcą się martwić o wysokie koszty rozwoju aplikacji.

4.2 Opis projektu

W tym rozdziale zostanie omówiony projekt autorski. Proces tworzenia projektu składał się z:

- stworzenia dwóch klasyfikatorów: Naiwnego klasyfikatora Bayesa, klasyfikatora oparty na sieciach neuronowych,
- przygotowania danych treningowych (rozdział 4.2.2),
- ewaluacji klasyfikatorów (rozdział 5),
- poprawienia danych treningowych (usunięcie szumów),
- ewaluacji klasyfikatorów po poprawieniu danych treningowych,
- wybraniu najlepszego klasyfikatora,
- stworzeniu interfejsu użytkownika.

[5] http://www.json.org/
[6] https://www.python.org/

4.2.1 Dane treningowe

Dane treningowe zostały stworzone, aby wytrenować stworzone klasyfikatory, które klasyfikują opis funduszu ze względu na jego kategorię oraz ze względu na region inwestowania. Dane treningowe składają się z:

- opisu polityki inwestycyjnej fundusz z dokumentu KIID[7],
- przynależności do klasy funduszu,
- przynależności do regionu inwestowania.

4.2.2 Naiwny klasyfikator Bayesa

Klasyfikator, opierający się na naiwnym algorytmie Bayesa, wykorzystuje zasadę działania opisaną w rozdziale 2.1. Klasyfikator został stworzony w języku Python.

Klasyfikator składa się z następujących klas:

- BagOfWords - tworzy tablicę słów występujących we wszystkich dokumentach, wraz z ich częstotliwością występowania,
- Document - jest odpowiedzialna za wczytywanie danych treningowych oraz ich normalizację[8],
- DocumentClass - jest odpowiedzialna za wyliczanie prawdopodobieństwa przynależności danego słowa do klasy,
- Model - tworzy model klasyfikatora oraz jest odpowiedzialna za klasyfikacje dokumentu do odpowiedniej klasy.

W pierwszym klasyfikatorze cechami są wszystkie słowa znajdujące się w BagOfWords, w drugim klasyfikatorze cechami są wybrane słowa z BagOfWords. Wybrane słowa to słowa wpływające na klasyfikację np. *akcje, obligacje, papiery wartościowe, Polska, Stany Zjednoczone*. Po wytrenowaniu modelu przez klasyfikator możemy wywołać funkcję *probability*, która zwraca tablicę prawdopodobieństw przynależności atrybutu (opis funduszu inwestycyjnego) do klas.

4.2.3 Klasyfikator oparty na sieciach neuronowych

Klasyfikator, opierający się na sieciach neuronowych, wykorzystuje zasadę działania opisaną w rozdziale 2.2. Do stworzenia sieci neuronowych została użyta biblioteka keras[9]. W pierwszym klasyfikatorze cechami są liczby występowań każdego słowa dostępnego w danych treningowych, a w drugim - cechami są liczby występowań wybranych słów ze wszystkich dostępnych w danych treningowych. Wybrane słowa to słowa wpływające na klasyfikację np. *akcje, obligacje, papiery wartościowe, Polska, Stany Zjednoczone*.

[7] https://www.analizy.pl/fundusze/temat-tygodnia/13990/kiid-%E2%80%93-klucz-ktory-otwiera-wiele-zamkow.html

[8] https://pl.wikipedia.org/wiki/Normalizacja_tekstu

[9] https://keras.io/

Cechy sieci neuronowych w obydwóch modelach:

- składają się z warstwy wejściowej, dwóch warstw ukrytych i warstwy wyjściowej,
- funkcjami aktywacji są funkcje *relu*,
- funkcja wyjścia jest funkcja $softmax$ określająca klasę.

Po wytrenowaniu modelu przez klasyfikator możemy wywołać funkcje *predict*, która zwraca tablicę prawdopodobieństw przynależności atrybutu (opis funduszu inwestycyjnego) do klas.

4.2.4 Baza danych

Baza danych została stworzona za pomocą biblioteki Sqlite[10]. Służy ona do przechowywania informacji na temat funduszy inwestycyjnych dostępnych na polskim rynku, które są pobierane przez program po klasyfikacji wprowadzonego przez użytkownika opisu funduszu. Baza danych zawiera tabelę, która składa się z trzynastu kolumn:

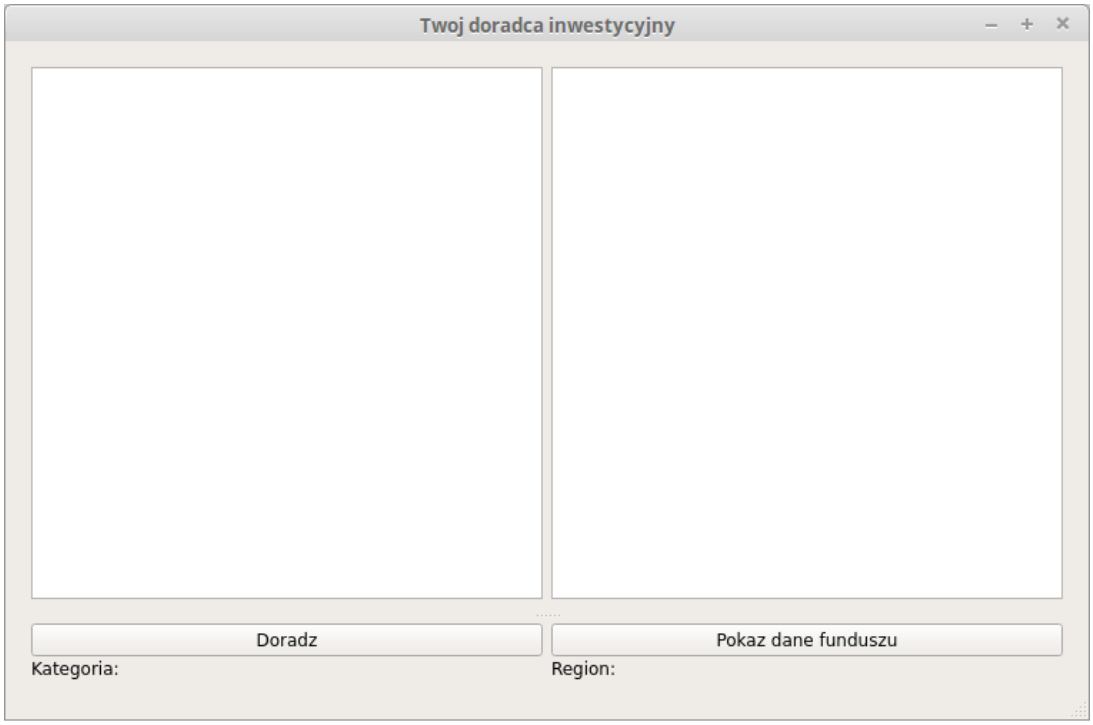
- Produkt - pełna nazwa funduszu,
- Kategoria - kategoria funduszu,
- Waluta - waluta funduszu,
- Region - region funduszu,
- Sektor - sektor funduszu,
- OplataManipulacyjna - opłata manipulacyjna funduszu,
- OplataStala - opłata za zarządzanie funduszu,
- OplataTER - rzeczywista opłata roczna za fundusz,
- PierwszaWplata - minimalna wartość pierwszej wpłaty,
- KIID - stopień ryzyka inwestycyjnego pobranego z dokumentu KIID,
- KartaProduktuLink - link do karty informacyjnej funduszu,
- KIIDLink - link do KIID funduszu,
- ProspektLink - link do prospektu funduszu.

Baza danych zawiera informacje o 990 funduszach inwestycyjnych.

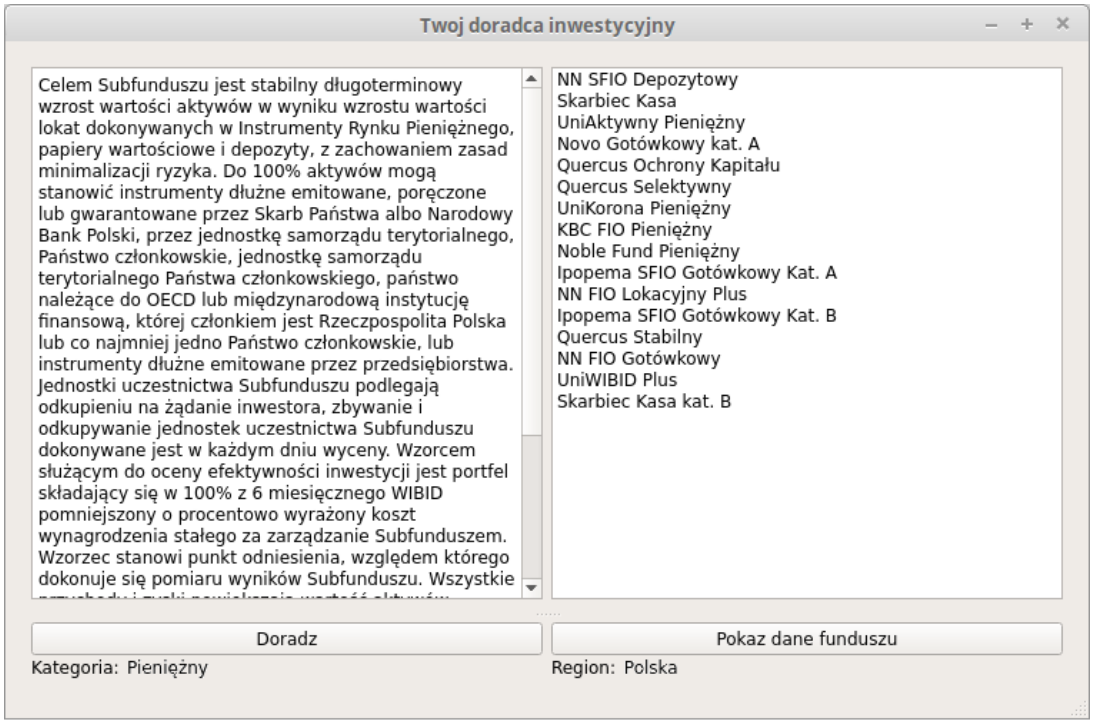
4.2.5 Interfejs użytkownika

Interfejs użytkownika został wykonany w języku Python z wykorzystaniem biblioteki PyQt5[11]. Interfejs użytkownika składa się z dwóch okien. W pierwszym oknie są dostępne 2 przyciski, pole tekstowe oraz pole widoku listy - rysunek 4.1.

[10] https://www.sqlite.org/
[11] http://pyqt.sourceforge.net/Docs/PyQt5/



Rysunek 4.1: Widok pierwszego okna



Rysunek 4.2: Widok pierwszego okna po wprowadzeniu opisu funduszu i użyciu przycisku "Doradź"

W polu tekstowym użytkownik umieszcza opis funduszu, który chce klasyfikować. Po naciśnięciu przycisku "Doradź" w polu widoku listy wyświetlane są nazwy funduszów o zbliżonej polityce inwestycyjnej do wprowadzonego przykładu. Przycisk "Doradź" służy do uruchamiania funkcji klasyfikującej wprowadzonego

tekstu oraz wprowadzenia wyników klasyfikacji do pól tekstowych znajdujących się poniżej przycisków. Przycisk "Pokaż dane funduszu" otwiera drugie okno, gdzie wyświetlane są informacje dotyczące zaznaczonego funduszu w polu widoku listy.

Przykładowy widok drugiego okna ukazany jest na rysunku 4.3. W oknie wyświetlane są informacje o funduszu takie jak:

- pełna nazwa funduszu,
- kategoria funduszu,
- waluta funduszu,
- region funduszu,
- sektor funduszu,
- opłata manipulacyjna funduszu,
- opłata za zarządzanie funduszu,
- rzeczywista opłata roczna za fundusz,
- minimalna wartość pierwszej wpłaty,
- stopień ryzyka inwestycyjnego pobranego z dokumentu KIID,
- link do karty informacyjnej funduszu,
- link do KIID funduszu,
- link do prospektu funduszu.

Wszystkie informacje są pobierane z bazy danych.

Rysunek 4.3: Widok drugiego okna po zaznaczeniu i użyciu przycisku "Pokaż dane funduszu"

4.2.6 Konserwacja i inżynieria wtórna

System korzysta z bazy danych pobierając informacje o aktualnych funduszach, dlatego też niezbędna jest aktualizacja informacji o funduszach przynajmniej raz na kwartał.

System klasyfikuje opis funduszu z wysoką dokładnością, lecz istnieje możliwość, że na rynku finansowym zmieni się schemat opisywania funduszy, co będzie skutkowało osłabieniem dokładności klasyfikatora. W takiej sytuacji będzie trzeba zaktualizować model klasyfikatora poprzez stworzenie nowych danych treningowych, lub stworzyć nowy klasyfikator. Rozwijać klasyfikator można na dwa sposoby:

- zwiększenie dokładności klasyfikatora poprzez modyfikację danych treningowych.
- zmianę modelu klasyfikatora poprzez utworzenie mniejszej liczby klas w celu otrzymywania bardziej zbliżonych funduszy inwestycyjnych z bazy danych do wprowadzanego opisu funduszu.

Rozdział 5

Ewaluacja i wyniki

Istnieje szereg empirycznych metod ewaluacji, niezależnych od klasyfikatora opartych na dokładności (z ang. accuracy) metod oceny algorytmu. Są nimi walidacja krzyżowa (z ang. cross-validation), jackknife[1] i bootstrap[2]. W niniejszej pracy do oceny algorytmów zostanie użyta walidacja krzyżowa prosta, walidacja krzyżowa 10-krotna oraz sprawdzenie dokładności na zbiorze testowym.

5.1 Walidacja krzyżowa

Walidacja krzyżowa zaczyna się od podzielenia zbioru treningowego T na n podzbiorów $(T_1, T_2, ..., T_n)$, przy czym każdy podzbiór nazywany jest próbą. Dla każdego n uczenie odbywa się na wszystkich próbach oprócz jednej:

$$T_s = T_1 \cup T_2 \cup \cdots \cup T_{n-1},$$

gdzie T_s to zbiór treningowy. Pominięty podzbiór jest wykorzystywany do oceny wybranego klasyfikatora C,

$$T_t = T_n,$$

gdzie T_t to zbiór testowy. Trening algorytmu i sprawdzanie dokładności obywa się dla wszystkich podzbiorów, które nie zostały użyte do oceny dokładności. Ze wszystkich ocen dokładności wyciągamy średnią i otrzymujemy ostateczną ocenę klasyfikatora.

[1] $https://en.wikipedia.org/wiki/Jackknife\_resampling$
[2] $https://en.wikipedia.org/wiki/Bootstrapping\_(statistics)$

Algorytm 1: Pseudokod algorytm walidacji krzyżowej

```
β, funkcja sprawdzająca poprawność klasyfikacji;
c, przydzielona klasa;
v, ilość poprawnie zaklasyfikowanych przykładów
Dane wejściowe: T, dane treningowe;
n, liczba;
a, testowany algorytm;
Wynik: Acc_a, dokładność algorytmu;
F[] = PodzielNaProby(T, n)
v ← 0
for i ← 1 to n do
    for j ← 1 to n do
        if i ≠ j then
            Trenuj(A, F[j])
        c ← WybierzKlase(a)
        v ← v + β(c, F[i])
Acc_a ← v/n
return Acc_a
```

5.2 Dane testowe

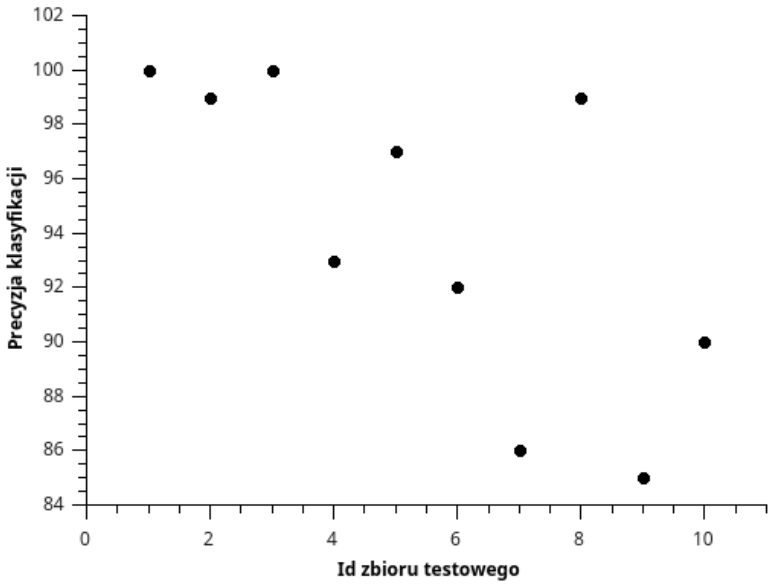
Zbiór testowy składa się z 100 opisów funduszów pobranych ze strony *analizy.pl*. Oprócz opisów zbiór testowy zawiera przynależność do kategorii funduszu oraz regionu inwestycji.

5.3 Ewaluacja naiwnego algorytmu Bayesa

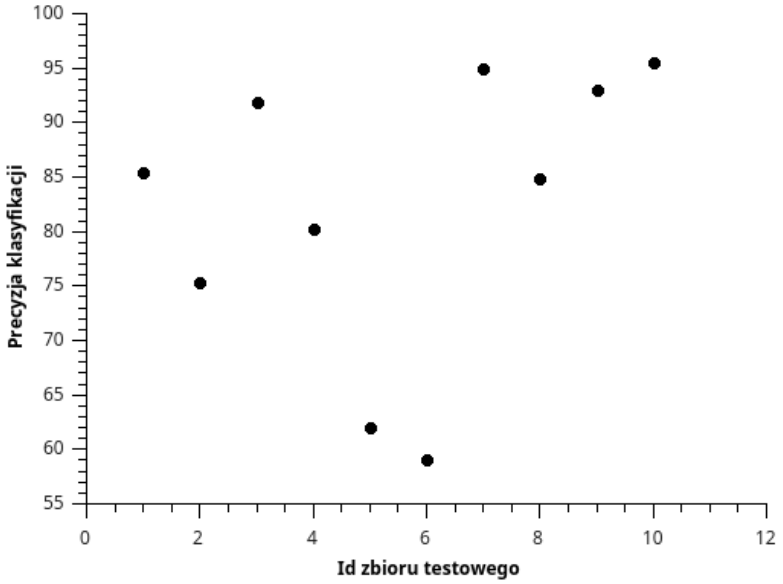
Ewaluacja na podstawie wszystkich słów znajdujących się w BagOfWords

W prostej walidacji krzyżowej zbiór treningowy został podzielony na: nowy zbiór treningowy składający się z 662 elementów i zbiór testowy składający się z 328 elementów. W tej walidacji klasyfikator określił kategorię funduszu z dokładnością na poziomie $94,20\%$, a region inwestycji funduszu z dokładnością na poziomie $85,67\%$.

W 10-krotnej walidacji krzyżowej zbiór treningowy został podzielony na 10 równych części. Każda część składała się z 99 elementów. Każda część tworzyła zbiór testowy, a reszta tworzyła zbiór treningowy. Na rysunku 5.3 widzimy z jaką dokładnością klasyfikator określił kategorię ze względu na kategorię funduszu. Średnia dokładność klasyfikatora ze względu na kategorię funduszu wynosiła $94,08\%$. Na rysunku 5.4 widzimy z jaką dokładnością klasyfikator określił kategorie ze względu na region inwestycji funduszu. Średnia dokładność klasyfikatora ze względu na region inwestycji funduszu wynosiła $82,22\%$.



Rysunek 5.1: Wykres dokładności klasyfikatora ze względu na kategorię funduszu dla każdego zbioru testowego



Rysunek 5.2: Wykres dokładności klasyfikatora ze względu na region inwestycji funduszu dla każdego zbioru testowego

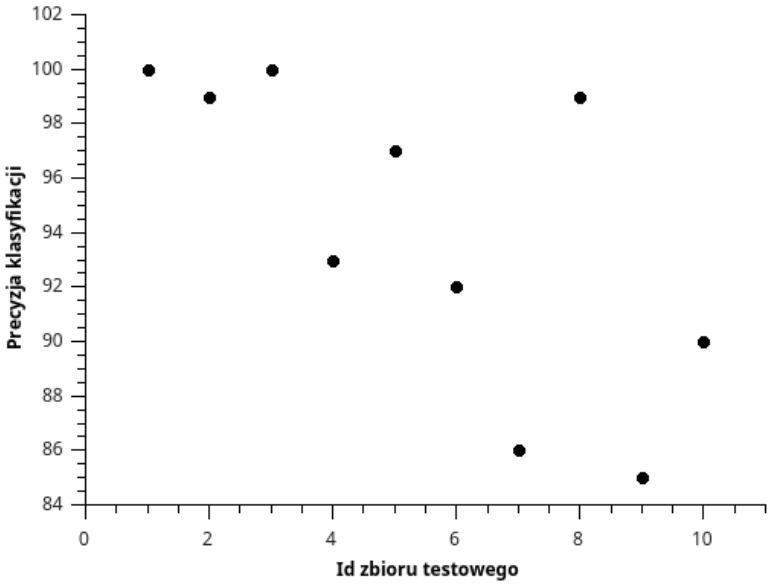
W walidacji na zbiorze testowym, klasyfikator określił kategorie funduszu z dokładnością na poziomie 61%, a region inwestycji funduszu z dokładnością na poziomie 67%.

Ewaluacja na podstawie wybranych słów znajdujących się w BagOfWords

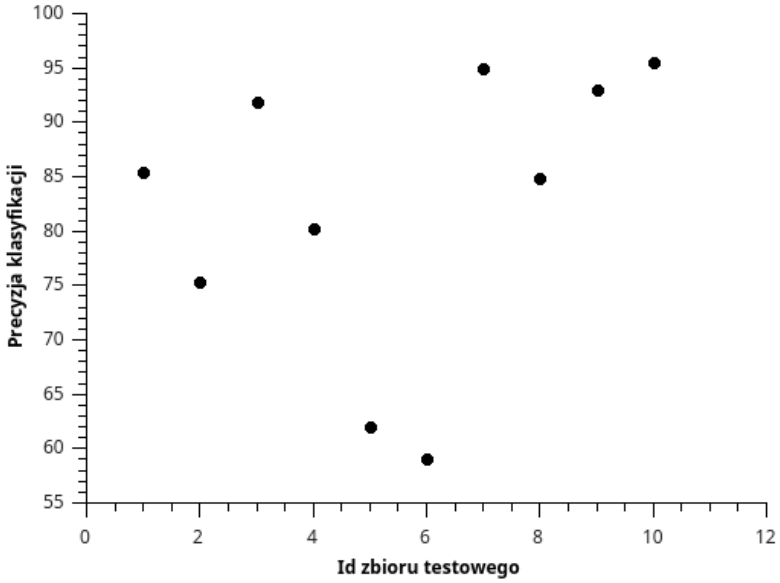
W prostej walidacji krzyżowej zbiór treningowy został podzielony na: nowy zbiór treningowy składający się z 662 elementów i zbiór testowy składający się z 328 elementów. W tej walidacji klasyfikator określił kategorie funduszu z dokładnością na poziomie $94,50\%$, a region inwestycji funduszu z dokładnością na poziomie $86,27\%$.

W 10-krotnej walidacji krzyżowej zbiór treningowy został podzielony na 10 równych części. Każda część składała się z 99 elementów. Każda część tworzyła zbiór testowy, a reszta tworzyła zbiór treningowy. Na rysunku 5.3 widzimy z jaką dokładnością klasyfikator określił kategorie ze względu na kategorię funduszu. Średnia dokładność klasyfikatora ze względu na kategorię funduszu wynosiła $93,98\%$. Na rysunku 5.4 widzimy z jaką dokładnością klasyfikator określił kategorie ze względu na region inwestycji funduszu. Średnia dokładność klasyfikatora ze względu na region inwestycji funduszu wynosiła $82,02\%$.

W walidacji na zbiorze testowym, klasyfikator określił kategorie funduszu z dokładnością na poziomie 62%, a region inwestycji funduszu z dokładnością na poziomie 66%.



Rysunek 5.3: Wykres dokładności klasyfikatora ze względu na kategorię funduszu dla każdego zbioru testowego



Rysunek 5.4: Wykres dokładności klasyfikatora ze względu na region inwestycji funduszu dla każdego zbioru testowego

5.4 Ewaluacja sieci neuronowych

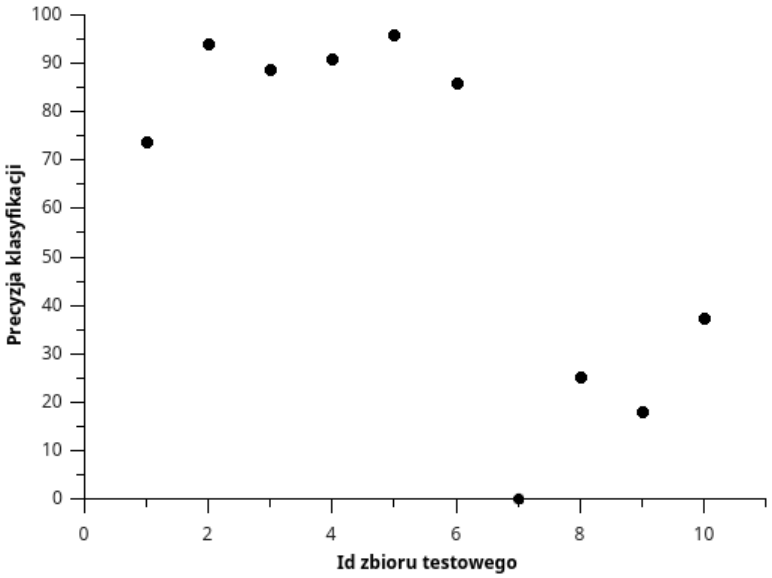
W klasyfikatorze wykorzystującym sieci neuronowe zastosowano takie same metody ewaluacji jak dla klasyfikatora wykorzystujący naiwny algorytm Bayesa w rozdziale 5.3.

Cechy sieci neuronowej jako liczby występowań każdego słowa dostępnego w danych treningowych

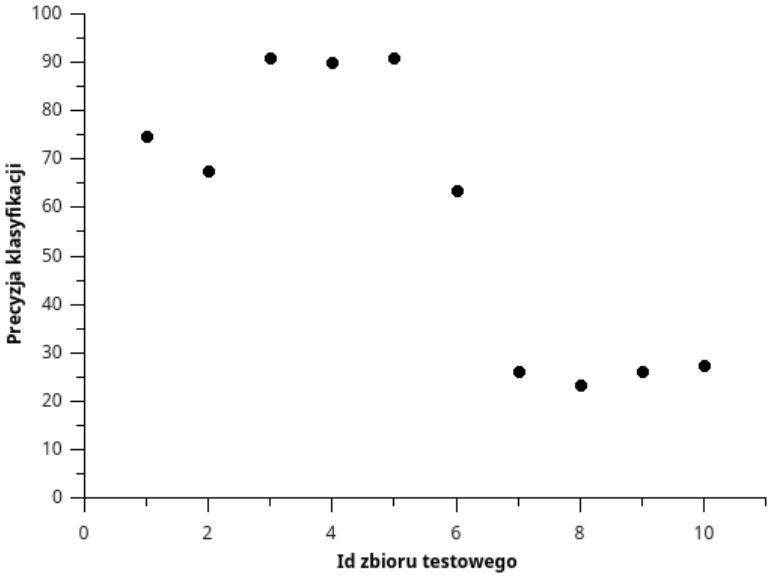
Dla prostej walidacji klasyfikator określił kategorie funduszu z dokładnością na poziomie $51,52\%$, a region inwestycji funduszu z dokładnością na poziomie $48,78\%$.

Dla 10-krotnej walidacji krzyżowej rysunek 5.5 przedstawia dokładność klasyfikatora ze względu na kategorię funduszu, a rysunek 5.6 ze względu na region inwestycji funduszu dla każdego utworzonego zbioru testowego. Średnia dokładność klasyfikatora ze względu na kategorię funduszu wynosiła $61,01\%$. Średnia dokładność klasyfikatora ze względu na region inwestycji funduszu wynosiła $58,08\%$.

W walidacji na zbiorze testowym klasyfikator określił kategorie funduszu z dokładnością na poziomie $53,00\%$, a region inwestycji funduszu z dokładnością na poziomie $54,00\%$.



Rysunek 5.5: Wykres dokładności klasyfikatora ze względu na kategorię funduszu dla każdego zbioru testowego



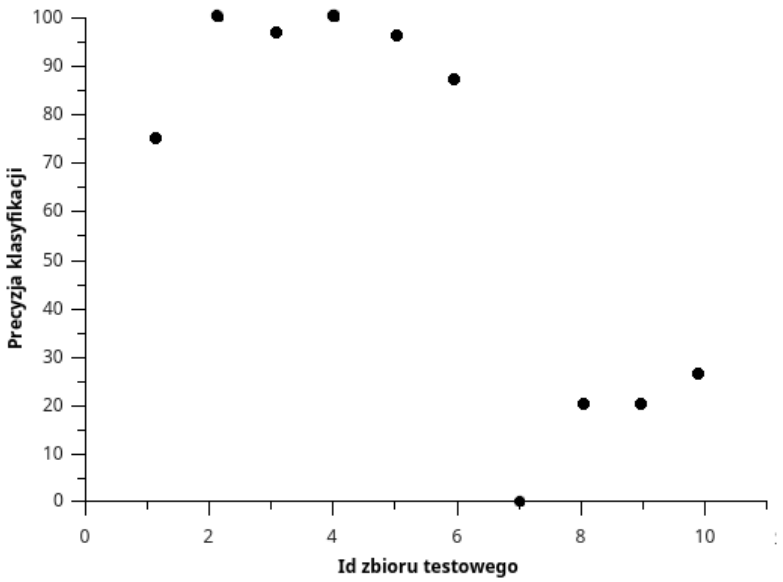
Rysunek 5.6: Wykres dokładności klasyfikatora ze względu na region inwestycji funduszu dla każdego zbioru testowego

Cechy sieci neuronowej jako liczby występowań wybranych słów

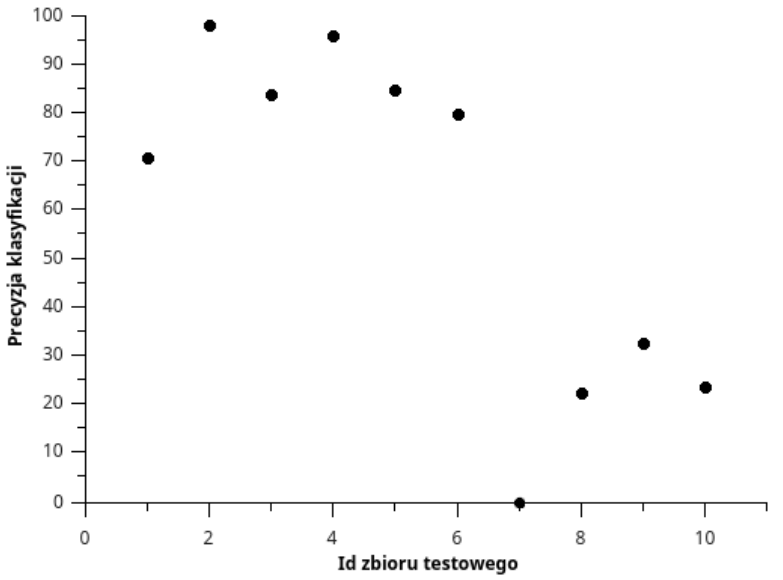
Dla prostej walidacji klasyfikator określił kategorie funduszu z dokładnością na poziomie $51,52\%$, a region inwestycji funduszu z dokładnością na poziomie $49,39\%$.

Dla 10-krotnej walidacji krzyżowej rysunek 5.7 przedstawia dokładność klasyfikatora ze względu na kategorię funduszu, a rysunek 5.8 ze względu na region inwestycji funduszu dla każdego utworzonego zbioru testowego. Średnia dokładność

klasyfikatora ze względu na kategorię funduszu wynosiła 52, 00%. Średnia dokładność klasyfikatora ze względu na region inwestycji funduszu wynosiła 58, 83%.



Rysunek 5.7: Wykres dokładności klasyfikatora ze względu na kategorię funduszu dla każdego zbioru testowego



Rysunek 5.8: Wykres dokładności klasyfikatora ze względu na kategorię funduszu dla każdego zbioru testowego

W walidacji na zbiorze testowym klasyfikator określił kategorie funduszu z dokładnością na poziomie 55, 00%, a region inwestycji funduszu z dokładnością na poziomie 57%.

5.5 Omówienie wyników

Wyniki ewaluacji klasyfikatorów przedstawia rysunek 5.1 i rysunek 5.2. Lepszym klasyfikatorem dla projektu "twój doradca inwestycyjny" okazał się NKB 1 wykorzystujący do klasyfikacji naiwny algorytm Bayesa. Na podstawie wyników można stwierdzić, że klasyfikator wykorzystujący naiwny algorytm Bayesa w porównaniu do klasyfikatora wykorzystującego sieci neuronowe wykazuje:

- krótszy czas uczenia się,
- większą tolerancje na szumy w danych treningowych,
- lepszą tolerancje na niewystarczającą ilość cech lub ich nadmiar,
- lepszą odporność na przetrenowanie (overfiting),

Klasyfikator wykorzystujący sieci neuronowe jest złożonym klasyfikatorem, który wymaga optymalizacji parametrów oraz wprowadzenia kryterium stopu uczenia się by zapobiec nadmiernemu dopasowaniu się sieci. Osiąga lepszą dokładność klasyfikacji po usunięciu mniej znaczących cech. Dla zbioru wykazuje nieznacznie gorszą dokładność klasyfikacji niż w klasyfikatorze wykorzystującym naiwny algorytm Bayesa.

	Prosta walidacja krzyżowa	**10-krotna walidacja krzyżowa**	**Walidacja na zbiorze testowym**
NKB 1	94,20	94,08	61,00
NKB 2	94,50	93,98	62,00
KOSN 1	51,25	61,00	53,00
KOSN 2	51,52	52,00	55,00

Tablica 5.1: Wyniki dokładności klasyfikacji ze względu na kategorię funduszu

	Prosta walidacja krzyżowa	**10-krotna walidacja krzyżowa**	**Walidacja na zbiorze testowym**
NKB 1	85,67	82,22	67,00
NKB 2	86,27	82,02	66,00
KOSN 1	48,78	58,00	54,00
KOSN 2	49,39	58,83	57,00

Tablica 5.2: Wyniki dokładności klasyfikacji ze względu na region inwestycji funduszu

5.6 Podsumowanie

Wybranym klasyfikatorem do projektu "twój doradca inwestycyjny" został klasyfikator wykorzystujący do klasyfikacji naiwny algorytm Bayesa. Dzięki wysokiej dokładności klasyfikacji projekt "twój doradca inwestycyjny" spełnia swoje założenia. Użytkownik po wprowadzeniu opisu funduszu otrzyma adekwatną listę funduszy inwestycyjnych, dzięki której będzie mógł zapoznać się ze wszystkimi możliwościami inwestycyjnymi na ryku odpowiadającymi jego profilowi inwestycyjnemu.

Bibliografia

[1] J. Schmidhuber, "Deep learning in neural networks: an overview", Neural Netw. 61, 85–117, 2014

[2] D. Lowd, P. Domingos, "Naive Bayes Models for Probability Estimation", ICML, 2005

[3] V. Sharma, S. Rai, A. Dev, "A Comprehensive Study of Artificial Neural Networks", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 10, 2012

[4] Oracle corporation, www.oracle.com, 2008

[5] Pablo D. Robles-Granda, Ivan V. Belik, "A Comparison of Machine Learning Classifiers Applied to Financial Datasets", World Congress on Engineering and Computer Science 2010 Vol I WCECS 2010, October 20-22, 2010

[6] Tom M. Mitchell, "Machine Learing", McGraw Hill, 1997

[7] Qiong Wang, George M. Garrity, James M. Tiedje, and James R. Cole1,"Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy",Appl Environ Microbiol, 2007

[8] S.L. Ting, W.H. Ip, Albert H.C. Tsang, "Is Naïve Bayes a Good Classifier for Document Classification?", International Journal of Software Engineering and Its Applications Vol. 5, No. 3, 2011

[9] Guobin Ou, Yi Lu Murphey, "Multi-class pattern classification using neural networks", ScienceDirect, Pattern Recognition 40 (2007) 4 – 18, 2016

Spis rysunków

Spis tablic

Spis algorytmów