AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

WYDZIAŁ ELEKTROTECHNIKI, AUTOMATYKI, INFORMATYKI I INŻYNIERII BIOMEDYCZNEJ

KATEDRA AUTOMATYKI I INŻYNIERII BIOMEDYCZNEJ

Praca dyplomowa magisterska

Klasyfikacja danych finansowych w celu wspomagania decyzji inwestycyjnych z wykorzystaniem klasyfikatora ASONN.

Classification of financial data in order to support investment decisions using classifier ASONN.

Autor: Andrzej Bobola
Kierunek studiów: Automatyka i Robotyka
Opiekun pracy: dr hab. Adrian Horzyk

Kraków, 2017

Uprzedzony o odpowiedzialności karnej na podstawie art. 115 ust. 1 i 2 ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (t.j. Dz.U. z 2006 r. Nr 90, poz. 631 z późn. zm.): „ Kto przywłaszcza sobie autorstwo albo wprowadza w błąd co do autorstwa całości lub części cudzego utworu albo artystycznego wykonania, podlega grzywnie, karze ograniczenia wolności albo pozbawienia wolności do lat 3. Tej samej karze podlega, kto rozpowszechnia bez podania nazwiska lub pseudonimu twórcy cudzy utwór w wersji oryginalnej albo w postaci opracowania, artystyczne wykonanie albo publicznie zniekształca taki utwór, artystyczne wykonanie, fonogram, wideogram lub nadanie.”, a także uprzedzony o odpowiedzialności dyscyplinarnej na podstawie art. 211 ust. 1 ustawy z dnia 27 lipca 2005 r. Prawo o szkolnictwie wyższym (t.j. Dz. U. z 2012 r. poz. 572, z późn. zm.) „Za naruszenie przepisów obowiązujących w uczelni oraz za czyny uchybiające godności studenta student ponosi odpowiedzialność dyscyplinarną przed komisją dyscyplinarną albo przed sądem koleżeńskim samorządu studenckiego, zwanym dalej „sądem koleżeńskim”, oświadczam, że niniejszą pracę dyplomową wykonałem osobiście i samodzielnie i że nie korzystałem ze źródeł innych niż wymienione w pracy.

Niniejszą pracę pragnę zadedykować mojej wspaniałej żonie, która wspierała i motywowała mnie w każdej chwili pisania pracy magisterskiej oraz rodzicom, dzięki którym miałem możliwość studiowania i zdobywania cennej wiedzy na Akademii Górniczo-Hutniczej oraz za wsparcie przez cały ten czas.

Pragnę także serdecznie podziękować promotorowi dr hab. Adrianowi Horzykowi za poświęcony czas oraz wsparcie podczas pisania pracy dyplomowej.

Spis treści

1 Wprowadzenie 5

2 System wspomagania decyzji inwestycyjnych 7

2.1 Rynki oraz dane finansowe 9

2.2 Proces wspomagania decyzji inwestycyjnej 13

2.3 Eksploracja danych i metody uczenia maszynowego 15

3 Klasyfikator ASONN 18

3.1 Konwersja danych do struktury AGDS 19

3.2 Budowa aktywnego asocjacyjnego grafu neuronowego AANG 20

3.2.1 Pierwotny proces tworzenia kombinacji 23

3.2.2 Modyfikacje procesu tworzenia kombinacji 25

3.3 Budowa asocjacyjnego klasyfikatora ASONN 26

4 Opis aplikacji 29

4.1 Model - implementacja ASONN 29

4.2 Dane, testy oraz symulacja 34

5 Testy 39

5.1 S&P 500 40

5.2 Kurs pary walutowej EUR/USD 44

5.3 Podsumowanie testów 47

6 Podsumowanie 49

7 Bibliografia 51

1 Wprowadzenie

W XXI wieku nastąpił gwałtowny rozwój technologiczny prowadzący do zwiększonej mocy obliczeniowej komputerów, która umożliwiła zastosowanie algorytmów komputerowych w wielu obszarach. Jednym z nich jest rynek finansowy, który obejmuje rynki: walutowy, kapitałowy oraz instrumentów pochodnych. W zależności od rynku inwestorzy używają odpowiednich narzędzi wspomagających ich proces podejmowania decyzji inwestycyjnych, należą do nich: analiza techniczna, analiza fundamentalna, metody statystyczne oraz techniki eksploracji danych. Głównym problemem dzisiejszego świata finansów jest znalezienie optymalnego sposobu do analizy danych finansowych, aby inwestorzy indywidualni oraz instytucjonalni byli w stanie efektywnie wykorzystać ogromną ilość informacji, która jest dostępna w dzisiejszych czasach w celu podjęcia decyzji inwestycyjnej.

W ostatnich latach bardzo dużym zainteresowaniem cieszą się badania naukowe dotyczące prognozowania wartości oraz kierunku poruszania się kursów walutowych oraz indeksów giełdowych przy wykorzystaniu metod eksploracji danych, które umożliwiają odnajdywanie wcześniej nieznanych wzorców, regularności oraz użytecznych informacji. Najczęstszym przedmiotem analizy naukowców były indeksy S&P 500 i DAX 30 oraz para walutowa EUR/USD. Według ankiety przeprowadzonej na Bond University z Australii [1], najbardziej popularnymi metodami były sztuczne sieci neuronowe oraz różne systemy hybrydowe, które zawierają w sobie wiele innych metod zarówno uczenia maszynowego jak i metod statystycznych.

W pracy skupiono się na klasyfikacji danych finansowych dotyczących indeksu S&P500 oraz pary walutowej EUR/USD. Klasyfikacja ma na celu zwrócenie informacji o przewidywanym kierunku poruszania się kursu na podstawie danych historycznych. Podejście to jest rezultatem wyników osiągniętych w pracy inżynierskiej autora, w której zastosowano sztuczną sieć neuronową typu multilayer perceptron w celu predykcji kursu pary walutowej EUR/USD, z których wynika, że można uzyskać znacznie lepszą skuteczność przy przewidywaniu kierunku przyszłego ruchu cen niż przy przewidywaniu wartości.

Celem niniejszej pracy jest stworzenie systemu wspomagania decyzji opartego na klasyfikatorze ASONN (asocjacyjna samo-optymalizująca się sieć neuronowa) w celu zbadania jego przydatności we wspomaganiu decyzji inwestycyjnych oraz porównania go z innymi systemami opartymi metodami uczenia maszynowego wymienionymi poniżej:

- Drzewa decyzyjne,
- K-NN (k najbliższych sąsiadów),
- SVM (maszyna wektorów nośnych),
- Sieci neuronowe.

Ponadto, celem autora pracy jest stworzenie aplikacji, która pozwoli na przejście całego procesu uczenia oraz testowania wyżej wymienionego klasyfikatora oraz wygenerowanie rzeczywistych prognoz inwestycyjnych wspomagających podejmowanie decyzji. Aplikacja posiada następujące funkcjonalności:

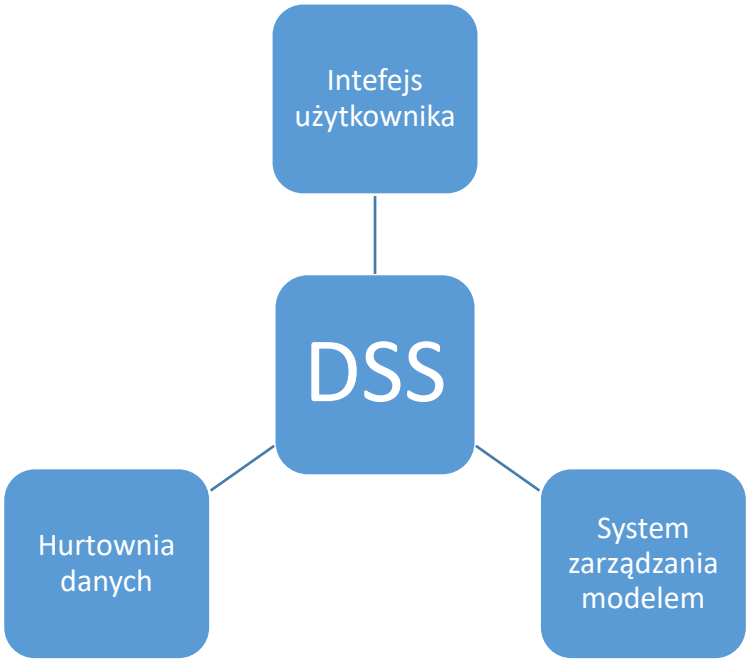
- Importowanie, definiowanie oraz przetwarzanie zbioru danych,
- Interfejs pozwalający na uczenie i testowanie klasyfikatora ASONN oraz innych metod uczenia maszynowego,
- Interfejs pozwalający na wygenerowanie decyzji inwestycyjnych oraz przetestowania systemu wspomagania decyzji na danych historycznych.

Do stworzenia aplikacji wykorzystano bazę danych PostgreSQL oraz PyCharm CE i język programowania Python. Klasyfikator ASONN został przystosowany do użytkowania z biblioteką Scikit-learn w Pythonie, co umożliwia wykorzystanie wielu metod związanych z uczeniem maszynowym.

Pozostała część pracy składa się z następujących rozdziałów. W pierwszym rozdziale zostanie przedstawiona struktura systemów wspomagających podejmowanie decyzji inwestycyjnych, następnie opisano przedmiot zainteresowania tej pracy, czyli inwestycje na rynkach finansowych oraz podejście badaczy do tematu klasyfikacji danych finansowych pod względem doboru metody klasyfikacji, doboru danych wejściowych oraz wyboru horyzontu czasowego. W kolejnym rozdziale opisano budowę, implementację oraz zasady działania klasyfikatora ASONN. Rozdział został oparty na pracy naukowej Pana dr hab. Adriana Horzyka pt. „Sztuczne systemy skojarzeniowe i asocjacyjna sztuczna inteligencja” [2]. W kolejnym rozdziale zostały opisane elementy systemu wspomagania decyzji stworzonego w ramach pracy magisterskiej. W ostatnim rozdziale zostały opisane testy wszystkich wyżej wymienionych metod oraz ich porównanie. Końcową częścią pracy jest podsumowanie uzyskanych wyników oraz pokazanie dalszych możliwości rozwoju klasyfikatora ASONN.

2 System wspomagania decyzji inwestycyjnych

Z definicji system wspomagania decyzji (Decision Support System) to skomputeryzowany system informacji używany w celu wspomagania procesu podejmowania decyzji w organizacji lub w przedsiębiorstwie. DSS pozwala na analizę bardzo dużej ilości danych i przedstawienie wiedzy w uproszczony sposób, co pozwala na szybsze rozwiązywanie problemów oraz ulepszenie podejmowanych decyzji w warunkach ciągłej zmienności, która jest spotykana przede wszystkim na rynkach finansowych. Duża zmienność oraz innowacje na rynkach finansowych spowodowały, że proces podejmowania decyzji w tych warunkach staje się niesamowicie skomplikowany i ryzykowny. Jednakże, jak wspomniano we wstępie, dzięki rozwojowi technologicznemu dostęp do rynku finansowego, danych finansowych oraz narzędzi informatycznych jest na najwyższym poziomie, co z kolei powoduje, że systemy wspomagania decyzji stają się niezbędne zarówno dla profesjonalistów jak i inwestorów indywidualnych w celu poprawienia efektywności inwestycyjnej. Typowa struktura systemu wspomagania decyzji wygląda następująco [3]:



Rysunek 2.1. Struktura systemu wspomagania decyzji

Hurtownia danych jest podstawowym elementem systemów wspomagania decyzji. Jest ona definiowana jako prezentująca wymiar czasowy, zintegrowana, nieulotna i przede wszystkim zorientowana tematycznie kolekcja danych lub jako kolekcja metod, technik i narzędzi

poprawiających zasoby informacyjne mające na celu wspieranie procesu podejmowania decyzji [4]. Jej zadaniem jest dostarczenie do programu wszystkich informacji i danych potrzebnych do zdefiniowania modelu oraz analizy. Natomiast system zarządzania modelem składa się ze wszystkich metod oraz technik analizy nieprzetworzonych danych potrzebnych do przeprowadzenia procesu decyzyjnego (analizy) oraz wygenerowania zrozumiałego wyniku dla odbiorcy końcowego, w przypadku tej pracy dla inwestora. Zarządzanie rozumiane jest jako ciągły rozwój oraz ulepszanie modelu w celu poprawiania skuteczności. Koniecznym elementem każdego systemu wspomagania decyzji jest interfejs użytkownika, który pozwala na wykorzystanie wszystkich możliwości modelu poprzez dostosowanie modyfikowalnych ustawień oraz na efektywną prezentację uzyskanych wyników [3].

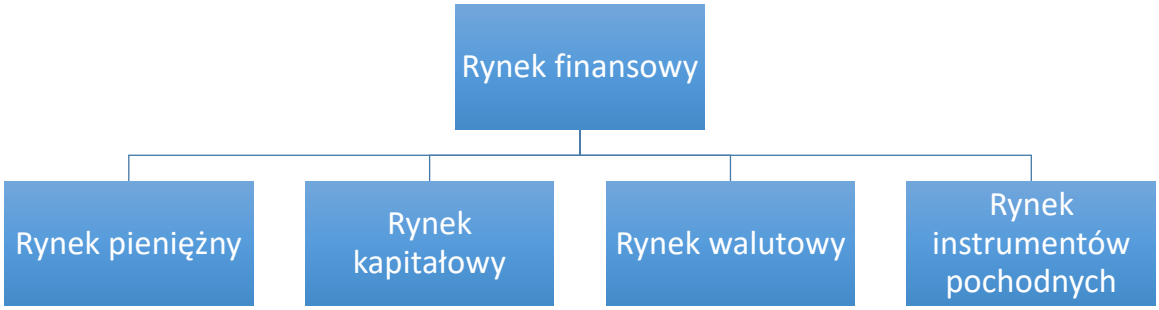
System wspomagania decyzji w celu dokonania udanej inwestycji powinien uwzględnić trzy podstawowe czynniki: prognozowanie cen, wyznaczanie odpowiedniego momentu zawarcia i zamknięcia transakcji oraz sposób zarządzania pieniędzmi [5]. Pierwszy z nich jest celem niniejszej pracy poprzez wykorzystanie klasyfikatora ASONN, który pozwala na określenie kierunku trendu rynkowego. Poprawne określenie kierunku jest koniecznym warunkiem udanej inwestycji. Kolejny czynnik jest równie ważny z uwagi na to, że mimo poprawnej oceny stanu danego rynku finansowego bez odpowiedniego wejścia i wyjścia z inwestycji możemy zanotować stratę. Natomiast uwzględnienie ostatniego czynnika pozwala na dywersyfikację ryzyka oraz minimalizowanie strat z inwestycji. Wszystkie te czynniki są zależne od rodzaju inwestycji. Inwestycje według ustawy o rachunkowości są to aktywa posiadane przez jednostkę w celu osiągania przychodu wynikającego z przyrostu wartości tych aktywów, odsetek, dywidend lub innych pożytków, a w szczególności aktywa finansowe, nieruchomości oraz wartości niematerialne i prawne, które są posiadane w celu osiągania tych korzyści [6]. Każda inwestycja posiada następujące cechy: czas realizacji, nakłady kapitałowe, oczekiwana stopa zwrotu oraz ryzyko realizacji inwestycji. Inwestycje można podzielić ze względu na długość trwania na długoterminowe (powyżej jednego roku) oraz krótkoterminowe (krócej niż jeden rok). W niniejszej pracy będą rozpatrywane inwestycje krótkoterminowe, na których inwestowanie często nazywane jest spekulacją ze względu na charakter tych inwestycji, której celem jest uzyskanie korzyści finansowej poprzez zmianę ceny instrumentu finansowego. Kolejnym ważnym czynnikiem dla procesu podejmowania decyzji inwestycyjnych są możliwości technologiczne oraz finansowe do analizy i monitorowania inwestycji (dostęp do informacji na temat otoczenia makroekonomicznego oraz sytuacji na rynkach finansowych). Czynniki te tworzą profil inwestora, do którego powinien być dostosowany system wspomagania decyzji inwestycyjnej.

2.1 Rynki oraz dane finansowe

Nawiązując do otoczenia makroekonomicznego inwestycji oraz sytuacji na rynkach finansowych, których wybór definiuje profil inwestora, konieczne jest podanie definicji rynku finansowego. Jest to miejsce, w którym zawierane są transakcje wymiany pieniądza lub instrumentów finansowych między podmiotami [7]. Spełnia on następujące funkcje:

- Alokacja kapitału – polega na przepływie środków finansowych pomiędzy podmiotami. Funkcja ta związana jest z kreowaniem rynku i przemieszczaniem kapitału między sektorami gospodarki i rynku finansowego, co sprzyja rozwojowi najbardziej dochodowych sektorów.
- Mobilizacja kapitału – polega na aktywnym tworzeniu procesów alokacji kapitału, co ma na celu zachęcić uczestników rynków finansowych do podejmowania decyzji dotyczących wprowadzania lub pozyskiwania kapitału z rynku.
- Wycena kapitału – polega na wycenie podmiotu, który został wprowadzony na rynek. Wycena związana jest z ceną instrumentu finansowego na podstawie, której inwestorzy oceniają atrakcyjność inwestycji.
- Transformacji kapitału – polega na szybkiej transformacji kapitału do innej formy w zależności od ryzyka i oczekiwanej stopy zwrotu z inwestycji.

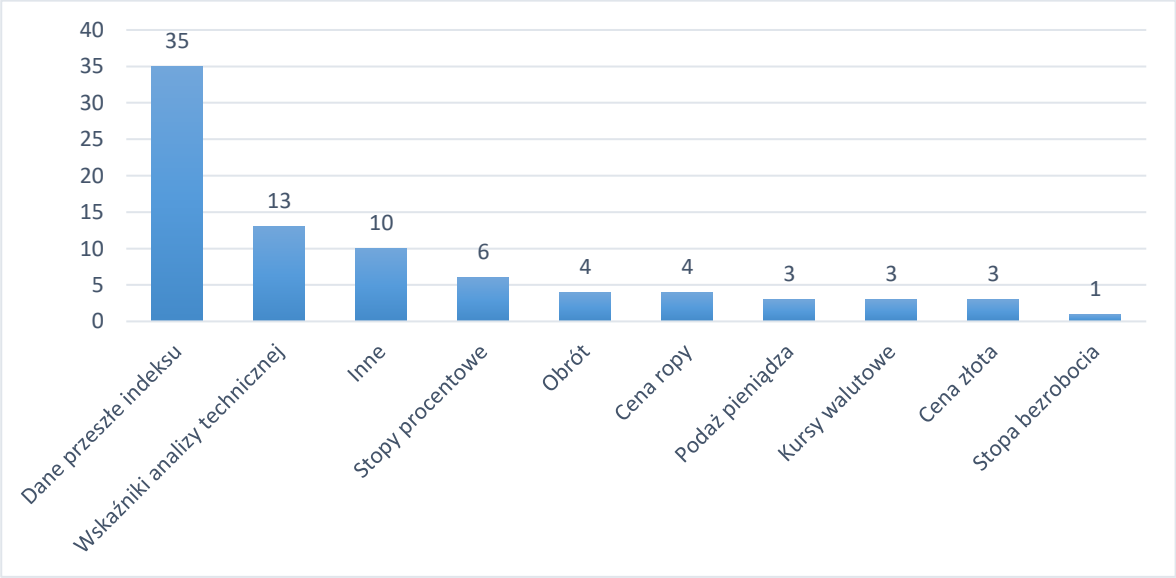
Wszystkie wyżej wymienione funkcję mają kluczowe znaczenie dla poprawnego funkcjonowania gospodarki oraz rynku finansowego. Na rysunku 2.2 przedstawiono strukturę rynku finansowego:



Rysunek 2.2. Rynek finansowy

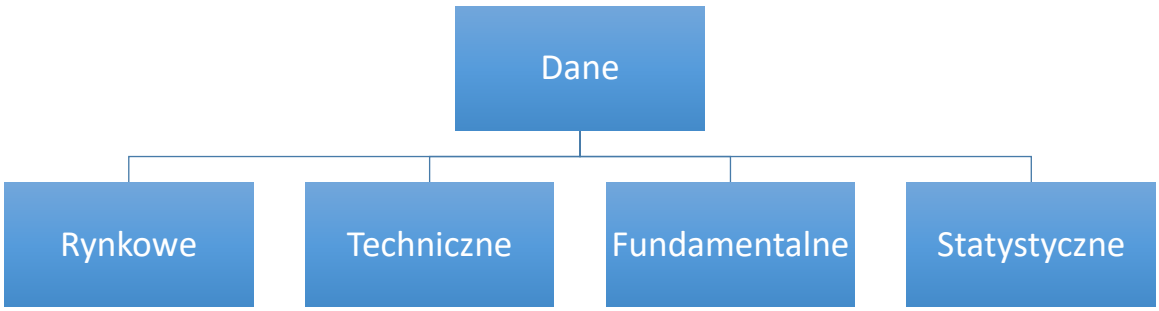
Rynek pieniężny jest to miejsce, w którym są zawierane transakcje o okresie zapadalności nie dłuższym niż jeden rok w formie gotówkowej lub bezgotówkowej. Do instrumentów rynku pieniężnego zalicza się: bony skarbowe, bony pieniężne, krótkoterminowe papiery dłużne, certyfikaty depozytowe, operacje otwartego rynku, lokaty międzybankowe, swapy walutowe, pożyczki i depozyty na rynku międzybankowym. Kolejnym segmentem rynku finansowego jest rynek kapitałowy, na którym są zawierane transakcje o okresie zapadalności dłuższym niż jeden rok. Jego rolą jest zaspokojenie potrzeb kapitałowych podmiotów gospodarczych w średnim i długim terminie. Jest on bardzo atrakcyjny dla inwestorów w celu inwestycji w papiery wartościowe o różnym poziomie ryzyka. Rynek walutowy (ang. forex) jest to miejsce, gdzie przedmiotem transakcji są waluty. Jest to rynek pozagiełdowy (Over-The-Counter), gdzie transakcje zawierane są na całym świecie niezależnie od pory dnia i nocy. Kursy walut ustalane są na podstawie skutków działania sił popytu i podaży. Głównymi uczestnikami rynku walutowego są banki, instytucje zbiorowego inwestowania, banki centralne dokonujące interwencji lub inwestycji rezerw, przedsiębiorstwa dokonujące zakupu lub sprzedaży walut lub zabezpieczające się przed niekorzystnymi zmianami kursów walutowych, a także osoby fizyczne [8]. Natomiast na rynku terminowym ma miejsce obrót instrumentami pochodnymi, których ceny zależą od cen aktywów podstawowych. Instrumenty pochodne, jakie funkcjonują w obrocie regulowanym to: kontrakty terminowe i opcje. Na polskim rynku finansowym najbardziej znanymi instrumentami pochodnymi są: kontrakty terminowe i opcje na indeksy giełdowe, akcje, obligacje, waluty, stopy procentowe.

W pracy skupiono się na najbardziej płynnych aktywach finansowych, tj. na rynku walutowym oraz na rynku instrumentów pochodnych ze względu na to, że są to główne miejsca spekulacji na rynkach finansowych.



Rysunek 2.3. Rodzaj danych najczęściej wybieranych przez badaczy [1]

Dotychczas przedstawiono strukturę oraz rolę rynków finansowych, natomiast najważniejszym aspektem z perspektywy systemów wspomagania decyzji są dane finansowe porządkowane w hurtowniach danych. Na rysunku 2.3 przedstawiono częstotliwość wyboru różnego rodzaju danych w pracach badawczych dotyczących analizy rynków finansowych. Dane te można przyporządkować do kategorii przedstawionych na rysunku 2.4.



Rysunek 2.4. Kategorie danych finansowych

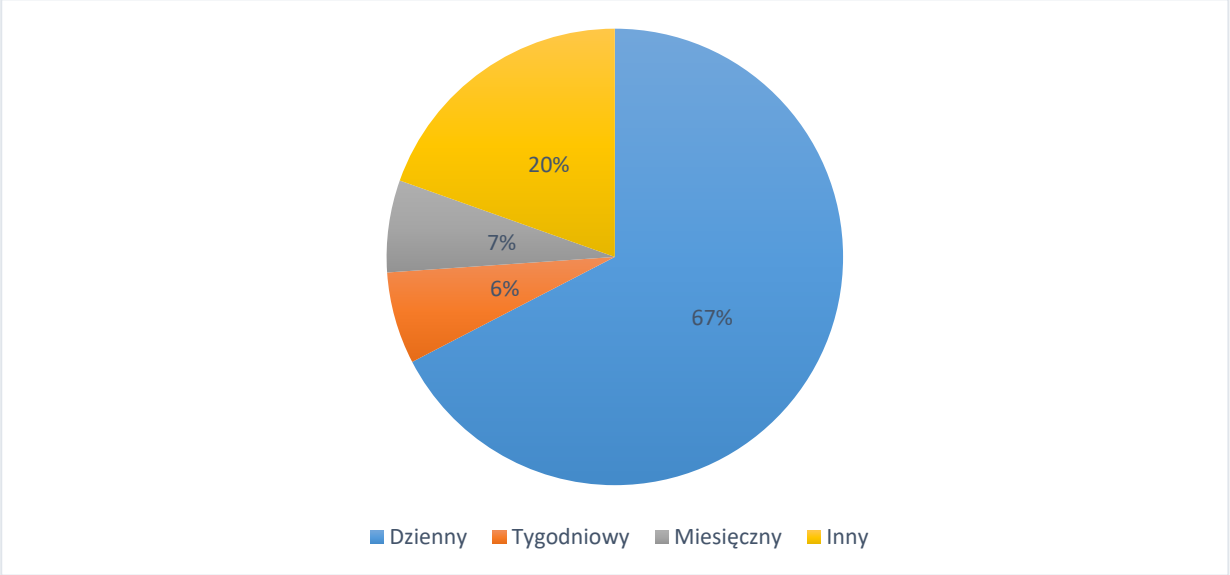
Głównym wyborem badaczy były przeszłe dane indeksów giełdowych, które zazwyczaj składają się ze spółek o najwyższej płynności oraz kapitalizacji. Są one wykorzystywane do określenia trendów oraz koniunktury panującej na danych rynkach oraz bardzo często służą jako punkt odniesienia do porównań strategii inwestycyjnych. Określane są jako dane rynkowe. Dane rynkowe są to dane pochodzące z notowań giełdowych różnego rodzaju instrumentów finansowych między innymi: kursy walut, notowania akcji, indeksów giełdowych oraz notowania instrumentów pochodnych. Najczęściej są one przedstawiane w formacie OHLC (open, high, low, close), który zawiera w sobie ceny otwarcia, najwyższą, najniższą oraz zamknięcia. Bardzo często do takich danych dodawany jest obrót, który wystąpił na danym rynku. Do tej kategorii zaliczamy również kursy walutowe oraz ceny surowców (złota i ropy). Dane te są podstawą do przeprowadzenia procesu decyzyjnego.

Na drugiej pozycji w ankiecie zostały sklasyfikowane wskaźniki analizy technicznej, które uzyskane są poprzez zastosowanie metod analizy technicznej na danych rynkowych. Zostały one stworzone na podstawie obserwacji zachowania ceny oraz wolumenu do pomocy w identyfikacji stanu rynku. Metody analizy technicznej są podstawą wielu strategii inwestycyjnych. Wyróżniamy wskaźniki: trendu, zmienności, impetu, siły rynku oraz wsparcia i oporu.

Kolejną z kategorii danych przedstawionych na rysunku 2.3 są dane fundamentalne, które opisują sytuację makroekonomiczną. Należą do nich: stopy procentowe, stopa bezrobocia oraz podaż pieniądza. Oprócz nich istnieje wiele innych, które są publikowane przez banki centralne, urzędy statystyczne oraz instytuty badawcze.

Ostatnią kategorią są dane statystyczne, do której należą dane uzyskane poprzez analizę statystyczną na danych rynkowych. Prawdopodobnie część tych danych została sklasyfikowana jako „Inne". Do tej kategorii zaliczamy: oczekiwane stopy zwrotu, miary zmienności (ryzyka: wariancja, odchylenie standardowe stopy zwrotu), miary wrażliwości (współczynnik beta) oraz korelacje między danymi.

Dane rynkowe oraz fundamentalne charakteryzują się częstotliwością występowania. W przypadku pierwszym, dane te mogą występować jako szeregi czasowe z częstotliwością mniejsza niż 1 sekunda. W drugim przypadku, dane te są zazwyczaj publikowane miesięcznie lub kwartalnie. Na rysunku 2.5 przedstawiono wybór badaczy dotyczący częstotliwości wybranych danych. Najbardziej popularne były dane dzienne (67%). Na drugim miejscu sklasyfikowano „Inne", których częstotliwość była większa niż dzienne. Najrzadziej wybierane były dane tygodniowe.



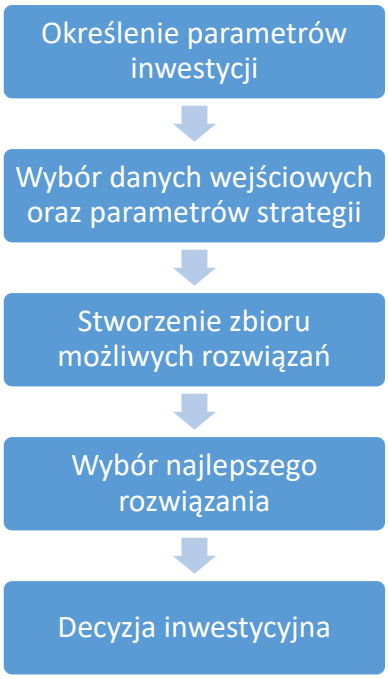
Rysunek 2.5. Częstotliwość najczęściej wybieranych danych przez badaczy

Biorąc pod uwagę podejście badaczy oraz własne doświadczenie do pracy wybrano następujące dane finansowe z dzienną częstotliwością (w przypadku danych fundamentalnych używane były najbardziej aktualne dane na dany dzień) [9] [10]:

1. Kurs pary walutowej EUR/USD,
2. Indeksy:
   - S&P 500 – w jego skład wchodzi 500 przedsiębiorstw notowanych na New York Stock Exchange i Nasdaq z najwyższą kapitalizacją.
   - DJI 30 – w jego skład wchodzi 30 największych przedsiębiorstw notowanych na New York Stock Exchange i Nasdaq
   - NASDAQ 100– w jego skład wchodzi 100 największych nie finansowych przedsiębiorstw notowanych na NASDAQ
   - DAX 30 – w jego skład wchodzi 30 największych firm notowanych na giełdzie niemieckiej DAX.
3. Surowce: złoto, srebro, pszenica, miedź, ropa naftowa (BRENT).
4. Wskaźniki analizy technicznej:
   - Wskaźniki Trendu: MA, EMA, WMA i BBANDS.
   - Wskaźniki impetu: ADX, CCI, DX, MACD, MFI, MOM, ROCP, RSI, STOCH i WILLR.
   - Wskaźniki zmienności: ATR i TRANGE
   - Wskaźniki siły rynku: AD, ADOSC i OBV.
5. Dane fundamentalne:
   - Stopy procentowe (Treasury Bonds w USA oraz Euribor w strefie Euro),
   - Inflacja (CPI – z perspektywy kupującego, PPI – z perspektywy sprzedającego),
   - Stopa bezrobocia,
   - Wskaźnik produkcji przemysłowej,
   - Wskaźnik zmian cen na rynku akcyjnym,
   - Wskaźnik sprzedaży detalicznej,
   - U.S. Crash Confidence Index – indeks pewności kryzysu,
   - U.S. Valuation Index – indeks wyceny rynków.
6. Dane statystyczne: oczekiwana stopa zwrotu, odchylenie standardowe, korelacje między danymi rynkowymi oraz współczynnik beta.

2.2 Proces wspomagania decyzji inwestycyjnej

Proces wspomagania decyzji inwestycyjnej w systemie zarządzania modelem polega na przetworzeniu danych w celu wygenerowania decyzji inwestycyjnej. Proces ten przedstawiono na poniższym rysunku.



Rysunek 2.6. Proces wspomagania decyzji inwestycyjnej

Określenie parametrów inwestycji dotyczy profilu inwestora, czyli wyboru rynku finansowego, horyzontu czasowego oraz wielkości inwestycji. Kolejnym krokiem jest wybór danych wejściowych do strategii inwestycyjnej (kombinacje wymienionych wcześniej danych) oraz zdefiniowanie początkowych parametrów strategii (możliwość zdefiniowania więcej niż jednej strategii). Po wybraniu danych konieczne jest ich przetworzenie (czyszczenie, integracja oraz selekcja) oraz transformacja do odpowiedniego formatu danych. Po zdefiniowaniu parametrów rozpoczyna się proces analizy danych oraz tworzenia możliwych decyzji inwestycyjnych, co jest celem niniejszej pracy. Następnie wybrane strategie są optymalizowane w celu wybrania optymalnego rozwiązania ze wszystkich strategii, bazując na skuteczności uzyskanej w testach, co pozwala na wygenerowanie ostatecznej decyzji inwestycyjnej.

Przechodząc do strategii podejmowania decyzji inwestycyjnych, najczęściej bazują one na poniższych metodach:

- Analiza techniczna,
- Analiza statystyczna,
- Analiza fundamentalna,
- Techniki eksploracji danych.

Z definicji analiza techniczna jest to badanie zachowania rynku, przede wszystkim przy użyciu wykresów, którego celem jest przewidywanie przyszłych trendów cenowych [11]. Jednakże,

dzięki możliwością technologicznym w dzisiejszych czasach coraz częściej analiza techniczna wykorzystywana jest przez komputery poprzez wykrywanie różnych zależności oraz wzorców na wykresach. Opiera się ona na trzech założeniach:

- Rynek dyskontuje wszystko, czyli wszystkie czynniki wpływające na rynkową cenę instrumentu finansowego znajdują odzwierciedlenie w cenie tego instrumentu.
- Ceny podlegają trendom, czyli ceny znajdują się w określonych trendach, które będą kontynuowane, dopóki nie nastąpi jego odwrócenie.
- Historia się powtarza, czyli analiza techniczna oparta jest na założeniu, że zachowania na rynku powtarzają się według określonych formacji cenowych, które sprawdziły się w przeszłości i zakłada się, że sprawdzą się w przyszłości.

Natomiast analiza statystyczna pozwala uzyskać informację o korelacjach, trendach i zależnościach pomiędzy różnymi danymi finansowymi za pomocą metod statystycznych oraz ekonometrycznych. Głównym obiektem badań w analizie statystycznej są modele ekonometryczne, które służą do objaśniania i prognozowania zmienności szeregów czasowych między innymi stóp procentowych, kursów walut oraz indeksów giełdowych [12]. Wyróżniamy tutaj między innymi modele klasy ARCH, GARCH oraz ARIMA.

Kolejnym wyżej wymienionym narzędziem jest analiza fundamentalna. Z definicji celem analizy fundamentalnej jest monitorowanie i klasyfikowanie aktywów finansowych pod względem ich jakości inwestycyjnej (jako szacunkowej oceny ryzyka) oraz oczekiwanej stopy zwrotu [13]. Polega ona na wyznaczeniu wartości wewnętrznej instrumentu i następnie podjęciu decyzji inwestycyjnej na podstawie aktualnych warunków ekonomicznych.

Eksploracja danych polega na odkrywaniu wiedzy w dużych bazach danych. Zadaniem metod eksploracji danych jest odnajdywanie nieznanych nietrywialnych zależności, trendów lub podobieństw, które ogólnie nazywane są wzorcami. Celem eksploracji jest analiza danych w taki sposób, aby uzyskać jak największą ilość informacji na dany temat w celu lepszego zrozumienia danego problemu.

2.3 Eksploracja danych i metody uczenia maszynowego

Eksploracja danych (ang. Data Mining) definiowana jest jako proces odnajdywania wzorców w danych. Jest ona jednym z etapów odkrywania wiedzy z baz danych (Knowledge Discovery in Databases), które w dzisiejszych czasach zawierają ogromne ilości informacji [14]. Eksploracja danych jest również definiowana jako proces analizy danych z różnych perspektyw prowadzącej do uzyskania użytecznej informacji [15]. Natomiast metody uczenia maszynowego (ang. Machine learning) są to metody stosowane między innymi w sztucznej

inteligencji w celu stworzenia automatycznego systemu potrafiącego uczyć się za pomocą dostarczanych danych oraz nabywania wiedzy na ich podstawie [16]. Przechodząc do połączenia eksploracji danych oraz danych finansowych, najczęściej wykorzystywane do analizy predykcyjnej są dwie techniki eksploracji danych: regresja oraz klasyfikacja. Są one używane do wielu zadań na przykład: oceny ryzyka kredytowego, wykrywanie nadużyć finansowych, segmentacja klientów oraz predykcja rynków finansowych. Główna różnica pomiędzy regresją, a klasyfikacją polega na tym, że regresja używana jest do predykcji wartości numerycznych lub ciągłych, natomiast klasyfikacja przyporządkowuje danym jedną ze zdefiniowanych kategorii, czyli służy do predykcji wartości dyskretnych lub nominalnych. W pracy skupiono się na technikach klasyfikacji.

Metoda klasyfikacji jest jedną z najważniejszych technik eksploracji danych z nadzorem, która ma bardzo szerokie zastosowanie w praktyce. Klasyfikacja składa się z następujących etapów: budowanie modelu, testowanie modelu oraz przyporządkowanie nowych danych do odpowiednich kategorii (klas, grup), co prowadzi do głównego celu klasyfikacji, czyli do zbudowania optymalnego modelu nazywanego klasyfikatorem. Pierwszym krokiem jest budowanie modelu na podstawie danych uczących, którego zadaniem jest odwzorowanie danych wejściowych na zbiór zdefiniowanych kategorii. Na podstawie modelu uzyskujemy wynik klasyfikacji dla danych testowych lub nowych obiektów dostarczanych do bazy danych.

W niniejszej pracy skupiono się na wykorzystaniu metod klasyfikacji we wspomaganiu decyzji inwestycyjnych na rynku finansowym, a przede wszystkim na porównaniu klasyfikatora ASONN z następującymi metodami uczenia maszynowego: drzewa decyzyjne, K-NN, sieci neuronowe oraz SVM. Wybór metod nastąpił na podstawie prac naukowych opisanych poniżej.

K najbliższych sąsiadów jest to jedna z najłatwiejszych metod uczenia maszynowego. Polega ona na wyliczaniu najczęstszej klasy obiektów występującej w danej odległości według wybranej miary odległości. K-NN została wybrana na podstawie pracy przewidującej wartość indeksu indyjskiej giełdy [17]. Autorzy tej pracy klasyfikowali dane z skutecznością 88% porównując ją z regresją logistyczną, przy której osiągnęli tylko 55% skuteczności.

Kolejną wybraną metodą była SVM, która uznawana jest za efektywną podczas przetwarzania dużej ilości danych. Jest w stanie klasyfikować więcej niż dwie klasy za pomocą funkcji liniowych lub nie-liniowych w zależności od jądra (kernel). Na tej podstawie wyznacza hiperpłaszczyznę zawierające wyniki klasyfikacji. W pracy przewidywano indeks giełdowy Hong Kongu. Autorzy dzięki odpowiedniemu przetwarzaniu danych oraz optymalizacji wektora danych wejściowych potrafili zwiększyć skuteczność klasyfikacji do 65%, co także jest bardzo dobrym wynikiem klasyfikacji w tak skomplikowanym zagadnieniu [18].

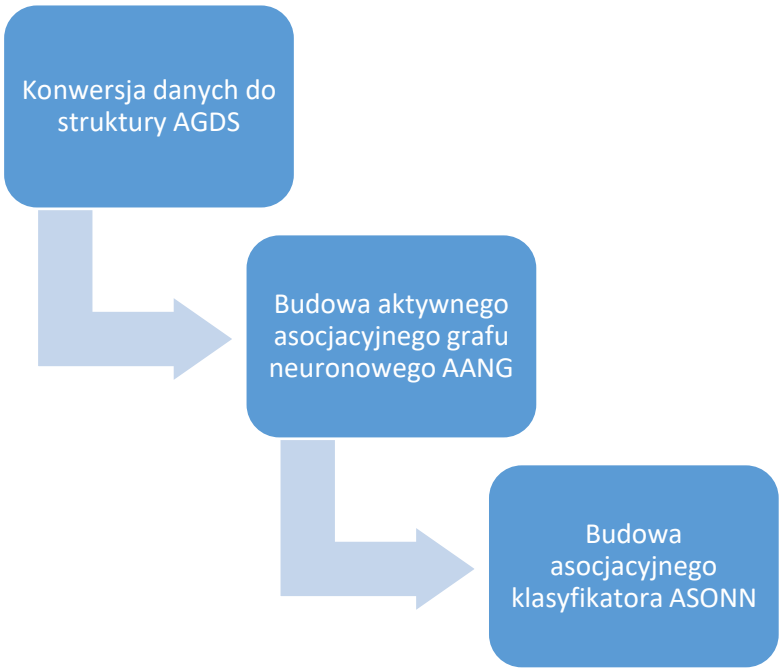
Drzewa decyzyjne są natomiast bardzo często używane ze względu na dużą łatwość w tworzeniu oraz interpretacji. Najpopularniejszą wersją tej metody jest C4.5, która jest znacznie bardziej odporna na szumy oraz pozwala na klasyfikację danych ciągłych. Ta metoda została

wybrana na podstawie badań nad predykcją australijskiej giełdy, gdzie uzyskano ponad 60% skuteczności dla tej metody [19].

Natomiast sieci neuronowe zostały wybrane na podstawie pracy inżynierskiej autora w celu ewaluacji podejścia z poprzedniej pracy. Sieci neuronowe są również jedną z najpopularniejszych metod ze względu na możliwość uczenia skomplikowanych zależności, które nie wątpliwie występują na rynkach finansowych [20].

3 Klasyfikator ASONN

Klasyfikator ASONN (asocjacyjna samo optymalizująca się sieć neuronowa) budowany jest na podstawie szybkiej analizy podobieństw wzorców uczących i wyznaczania dyskryminatywnych kombinacji cech ich definiujących. Budowa klasyfikatora polega na zbudowaniu reprezentatywnych kombinacji, które dążą do reprezentacji wszystkich wzorców uczących, dzięki czemu uzyskuje ona bardzo wysoką skuteczność, a zastosowanie rozmytych przedziałów wartości prowadzi do uzyskania bardzo dobrych wyników uogólniania. Sposób tworzenia kombinacji może być modyfikowany poprzez zastosowanie wiedzy eksperckiej do poszczególnych cech lub minimalizacji kosztu tworzenia klasyfikatora poprzez dobór odpowiednich wag dla droższych atrybutów. Klasyfikator nie wymaga wstępnego przetwarzania danych oraz definiowania żadnych parametrów poza opcjonalnymi wartości dotyczącymi kosztu lub wiedzy eksperckiej. Obsługuje on zarówno wartości numeryczne jak i symboliczne, dla których stosowane jest odmienne podejście. W implementacji zastosowano znaczne uproszczenia dotyczące odwzorowania biologicznych procesów odbywających się w neuronach, co pozwala na znaczne przyspieszenie działania sieci przy niepogorszonym działaniu. W następujących podrozdziałach opisano poszczególne etapy tworzenia klasyfikatora ASONN, które zostały przedstawione na Rys. 3.1 oraz implementacje klasyfikatora i oprogramowania potrzebnego do przeprowadzenia badań dotyczących przydatności użytkowania ASONN na rynkach finansowych.



Rysunek 3.1. Etapy budowy klasyfikatora ASONN.

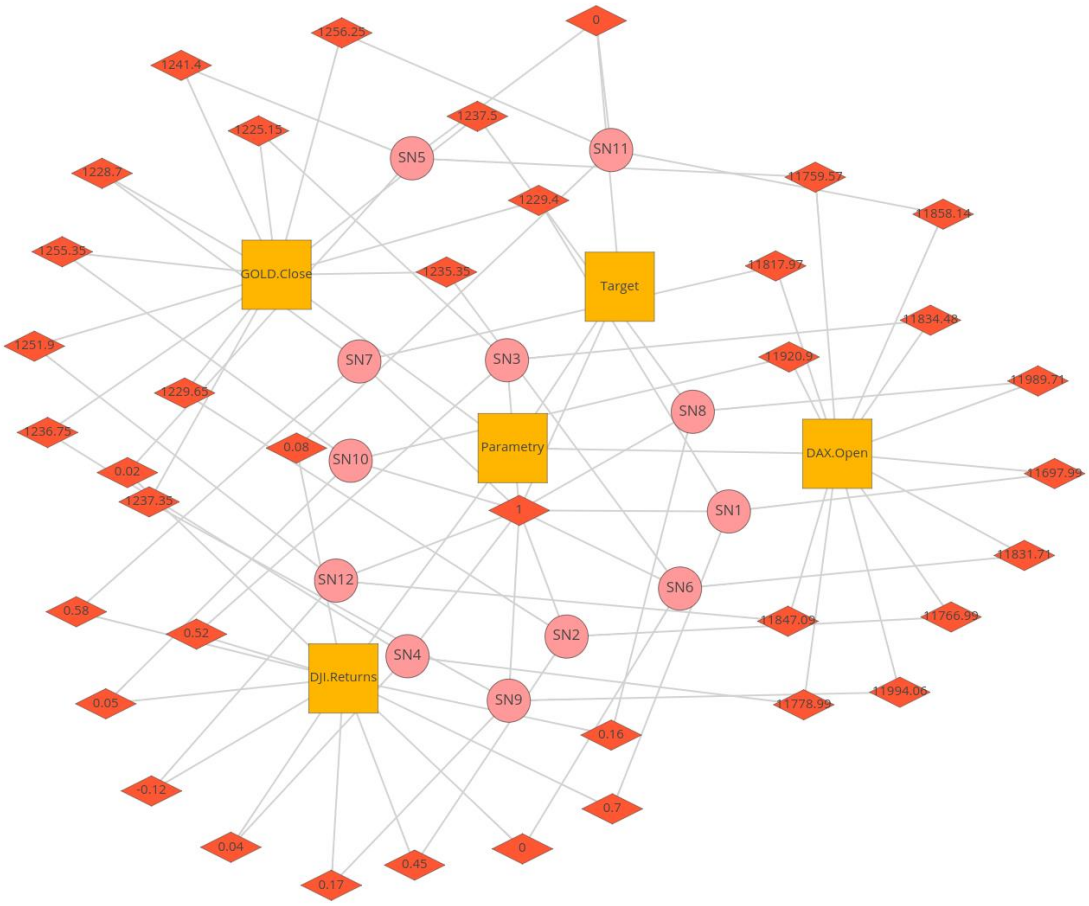
3.1 Konwersja danych do struktury AGDS

Pierwszym etapem budowy klasyfikatora ASONN jest utworzenie asocjacyjnej grafowej struktury danych poprzez konwersję danych przechowywanych w tabelach w relacyjnych bazach danych. Jest ona przeprowadzana w taki sposób, że nazwy atrybutów, nazwy rekordów (klucz) oraz każde pola rekordów tworzą nowe węzły. Następnie węzły, które reprezentują pola rekordów są łączone krawędziami EDEF z węzłami reprezentującymi nazwy tych rekordów oraz z odpowiadającymi im węzłami atrybutów. W przypadku występowania duplikatów rekordów zwiększana jest wartość węzła, który reprezentuje dany duplikat. Natomiast w przypadku występowania duplikatów pól rekordów nie jest tworzony nowy węzeł, tylko dodawane jest odpowiednie połączenie krawędzią EDEF. Kolejnym etapem jest połączenie węzłów reprezentujących wartości atrybutów krawędziami ESIM w przypadku, gdy istnieje relacja porządkująca w zbiorze tych wartości.

W ten sposób uzyskuje się niezduplikowane oraz posortowane listy dla poszczególnych atrybutów. Dzięki takiej konwersji z AGDS możemy wyznaczyć duplikaty, wzorce sprzeczne (dwie krawędzie EDEF do węzłów nazw dla takich samych węzłów wartości), podobieństwa oraz korelacje, które będą potrzebne w dalszych etapach budowy klasyfikatora ASONN. Etap ten został przedstawiony na podstawie tabeli 3.1 oraz rysunku 3.2

Tabela 3.1. Dane uczące do prezentacji procesu budowanie klasyfikatora ASONN

ID	GOLD.Close	DAX.Open	DJI.Close.Returns	Target
1	1229.4	11697.99	0.7	1
2	1229.65	11766.99	0.45	1
3	1225.15	11834.48	0.52	1
4	1236.75	11778.99	0.04	1
5	1241.4	11759.57	0.02	0
6	1235.35	11831.71	0.0	1
7	1228.7	11817.97	0.58	1
8	1237.5	11989.71	0.16	1
9	1237.35	11994.06	0.17	1
10	1255.35	11920.9	0.05	1
11	1256.25	11858.14	0.08	0
12	1251.9	11847.09	-0.12	1



Rysunek 3.2. Graf AGDS po konwersji z tabeli 3.1

3.2 Budowa aktywnego asocjacyjnego grafu neuronowego AANG

Kolejnym etapie budowy klasyfikatora ASONN jest budowa aktywnego asocjacyjnego grafu neuronowego. Istnieją dwie możliwości budowy takiego grafu:

1. Pierwsze podejście polega na wykorzystaniu struktury AGDS, która następnie przekształcana jest do grafu AANG poprzez zamianę węzłów na neurony oraz krawędzi na ważone połączenia.
2. Drugie podejście polega na wykorzystaniu mechanizmu asocjacyjnego sortownia ASSORT. Polega on na zbudowaniu inicjalnej struktury grafu AANG, która składa się z SENSINów, a następnie na pobudzaniu grafu AANG i stopniowej rozbudowie o kolejne wzorce.

W pracy skupiono się na pierwszym podejściu ze względu na prostotę i szybsze działanie w porównaniu do drugiego podejścia. Węzły atrybutów zostały zamienione na SENSINy (wejścia sensoryczne), węzły wartości wzorców na VN (receptoryczny neuron wartości), węzły wzorców na SN (neuron wzorca) oraz węzły nazwy rekordów na CN (neuron klasy). Krawędzie EDEF zostały zamienione na połączenia ADEF (asocjacyjna definicja) oraz krawędzie ESIM na połączenia ASIM (asocjacyjne podobieństwo). Kolejnym krokiem jest wyznaczenie wartości połączeń, które przyjmują następujące wartości:

- Dla połączenia ASIM:

$$w_{ASIM}^{P} = \left(\frac{Range^{P} - |Val - Val_{ASIM}|}{Range^{P}}\right)^{n} \tag{3.1}$$

$$Range^{P} = Max^{P} - Min^{P} \tag{3.2}$$

- Dla połączenia ADEF pomiędzy VN, a SN:

$$w_{VN,SN}^{ADEF} = \frac{\frac{\|\{\overline{SN}:VN \leftrightsquigarrow \overline{SN} \leftrightsquigarrow CN\}\|}{\|\{\overline{SN}:VN \leftrightsquigarrow \overline{SN}\}\|}}{\sum_{\overline{VN}_i \leftrightsquigarrow SN} \frac{\|\{\overline{SN}:\overline{VN}_i \leftrightsquigarrow \overline{SN} \leftrightsquigarrow CN\}\|}{\|\{\overline{SN}:\overline{VN}_i \leftrightsquigarrow \overline{SN}\}\|}} \tag{3.3}$$

- Dla połączeń ADEF pomiędzy SN, a VN i CN:

$$w_{SN,VN}^{ADEF} = 1 \tag{3.4}$$

$$w_{SN,CN}^{ADEF} = 1 \tag{3.5}$$

- Dla połączenia ADEF pomiędzy CN, a SN:

$$w_{CN,SN}^{ADEF} = 1 \tag{3.6}$$

Próg aktywacji wyżej wymienionych neuronów jest zawsze równy 1.

Przed rozpoczęciem procesu tworzenia kombinacji konieczne jest wyznaczenie na podstawie wcześniej uzyskanej struktury grafu neuronowego AANG korelacji własnych oraz korelacji obcych neuronów wzorców. Korelacje własne są to korelacje pozytywne w kontekście tworzenia kombinacji. Występują one w momencie, gdy dwa dowolne neurony SN są

połączone z tym samem neuronem CN. W przeciwnym wypadku mamy do czynienia z korelacjami obcymi, które wpływają negatywnie na proces tworzenia kombinacji z powodu pogorszenia dyskryminatywności oraz reprezentatywności kombinacji.

Kolejnym etapem jest tworzenie neuronów kombinacji (KN). Proces tworzenia kombinacji rozpoczynamy od najbardziej negatywnie skorelowanego neuronu wzorca SN, następnie do kombinacji dodajemy kolejne neurony SN tak długo, dopóki kombinacja nie osiągnie odpowiedniej reprezentatywności przy określonym stopniu dyskryminatywności oraz określonym sposobie tworzenia neuronów kombinacji. Następne neurony KN są tworzone analogicznie wykorzystując kolejne najbardziej negatywnie skorelowane neurony SN. Częścią procesu tworzenia neuronów kombinacji jest tworzenie neuronów przedziału RN oraz neuronów podzbiorów UN. Ten sposób tworzenia kombinacji umożliwia uzyskanie redundancji poprzez możliwość połączenia z różnymi neuronami kombinacji. W niniejszej pracy zaimplementowano sposób tworzenia kombinacji, który został przedstawiony w książce dr Adriana Horzyka [2] oraz zaproponowano trzy zmodyfikowane sposoby tworzenia kombinacji. W celu opisu metod tworzenia kombinacji konieczne jest zdefiniowanie następujących pojęć:

- Stopień dyskryminacji – służy do określenia ilości cech na podstawie, których możemy rozróżnić przeciwne wzorce:
  $$maxDiscrim_{KN} = \|\{RN: RN \nleftrightarrow KN \nleftrightarrow CN\} \cup \{UN: UN \nleftrightarrow KN \nleftrightarrow CN\}\| - \max_{outSN \nleftrightarrow CN} \|\{outSN \nleftrightarrow VN \nleftrightarrow germSN\}\| \quad (3.7)$$

- Współczynnik „Weeds”- służy do określenia negatywnego wpływu wzorców obcych na poszerzanie kombinacji za pomocą odpowiednich neuronów przedziału lub podzbiorów. Jest on wyznaczany według poniższych wzorów:

  $$Weeds_{RN} = \sum_{outSN \nleftrightarrow CN \nleftrightarrow KN} \|\{VN \nleftrightarrow RN: VN \nleftrightarrow outSN\}\|^2 \quad (3.8)$$

  $$Weeds_{UN} = \sum_{outSN \nleftrightarrow CN \nleftrightarrow KN} \|\{VN \nleftrightarrow UN: VN \nleftrightarrow outSN\}\|^2 \quad (3.9)$$

  $$AllWeeds_{KN} = \sum_{RN \nleftrightarrow KN} Weeds_{RN} + \sum_{UN \nleftrightarrow KN} Weeds_{UN} \quad (3.10)$$

- Współczynnik „Seeds” - służy do określenia pozytywnego wpływu wzorców własnych na poszerzanie kombinacji za pomocą odpowiednich neuronów przedziału lub podzbiorów. Jest on wyznaczany na podstawie poniższych wzorów:

  $$Seeds_{RN} = \sum_{inSN \nleftrightarrow CN \nleftrightarrow KN} \|\{VN \nleftrightarrow RN: VN \nleftrightarrow inSN\}\|^2 \quad (3.11)$$

$$Seeds_{UN} = \sum_{inSN \leftrightsquigarrow CN \leftrightarrow KN} \|\{VN \leftrightsquigarrow UN : VN \leftrightsquigarrow inSN\}\|^2 \quad (3.12)$$

$$AllSeeds_{KN} = \sum_{RN \leftrightsquigarrow KN} Seeds_{RN} + \sum_{UN \leftrightsquigarrow KN} Seeds_{UN} \quad (3.13)$$

3.2.1 Pierwotny proces tworzenia kombinacji

W tym podrozdziale opisano proces tworzenia kombinacji na podstawie książki „Sztuczne systemy skojarzeniowe i asocjacyjna sztuczna inteligencja" [2]. Pierwszym krokiem jest wyznaczenie wymienionych poniżej współczynników (3.14 oraz 3.15) określających liczbowo opłacalność poszerzania w danym kierunku dla każdego neuronu przedziału oraz podzbioru. Celem jest wybranie współczynnika o maksymalnej wartości.

Dla neuronów RN:

$$dir^{-}_{P,KN} = \frac{1}{\gamma_p} \cdot \sum_{VN^-} dir^{VN^-}_{P,KN} \quad (3.14)$$

$$dir^{+}_{P,KN} = \frac{1}{\gamma_p} \cdot \sum_{VN^-} dir^{VN^+}_{P,KN} \quad (3.15)$$

$$dir^{VN^-}_{P,KN} = \delta^{distmin}_{P,KN,VN^-} \cdot \begin{pmatrix} \sum_{\substack{inSN \leftrightsquigarrow VN^- \wedge \\ inSN \leftrightsquigarrow CN \leftrightsquigarrow KN}} \delta^{repr}_{inSN} \cdot \delta^{cont}_{inSN,KN} \\ - \sum_{\substack{outSN \leftrightsquigarrow VN^- \wedge \\ outSN \leftrightarrow KN}} \delta^{cont}_{outSN,KN} \end{pmatrix} \quad (3.16)$$

$$dir^{VN^+}_{P,KN} = \delta^{distmax}_{P,KN,VN^+} \cdot \begin{pmatrix} \sum_{\substack{inSN \leftrightsquigarrow VN^+ \wedge \\ inSN \leftrightsquigarrow CN \leftrightsquigarrow KN}} \delta^{repr}_{inSN} \cdot \delta^{cont}_{inSN,KN} \\ - \sum_{\substack{outSN \leftrightsquigarrow VN^+ \wedge \\ outSN \leftrightarrow KN}} \delta^{cont}_{outSN,KN} \end{pmatrix} \quad (3.17)$$

Dla neuronów UN:

$$dir^{VN_i}_{P,KN} = \frac{1}{\gamma_p} \cdot \begin{pmatrix} \sum_{\substack{inSN \leftrightsquigarrow VN_i \wedge \\ inSN \leftrightarrow KN}} \delta^{repr}_{inSN} \cdot \delta^{cont}_{inSN,KN} \\ - \sum_{\substack{outSN \leftrightsquigarrow VN_i \wedge \\ outSN \leftrightarrow KN}} \delta^{cont}_{outSN,KN} \end{pmatrix} \quad (3.18)$$

Gdzie γ_p symbolizuje koszt uzyskania danego atrybutu (w przypadku odpłatnych danych) lub umożliwia zastosowanie wiedzy eksperckiej w celu zdefiniowania przydatności danego atrybutu.

Aby obliczyć wartości współczynników konieczne jest obliczenie wszystkich składowych powyższych wzorów:

- Współczynniki distmin oraz distmax służą do określenia odległości VN od wartości odpowiednio minimalnej oraz maksymalnej neuronu RN w stosunku do przedziału wartości VN w kombinacji KN.

$$\delta_{P,KN,VN}^{distmin} = \left(1 - \frac{|RN.\min - VN.val|}{Range_P^{KN}}\right)^2 \quad (3.19)$$

$$\delta_{P,KN,VN}^{distmax} = \left(1 - \frac{|RN.\max - VN.val|}{Range_P^{KN}}\right)^2 \quad (3.20)$$

gdzie:

$$RN.\min > VN^{-}.val \in P \quad (3.21)$$

$$RN.\max > VN^{+}.val \in P \quad (3.22)$$

$$\text{Range}_P^{KN} = Max_P^{KN} - Min_P^{KN} \quad (3.23)$$

$$Max_P^{KN} = \max_{VN.val \in P}\{VN.val: \exists_{SN \leftrightsquigarrow CN \leftrightsquigarrow KN} VN \leftrightsquigarrow SN\} \quad (3.24)$$

$$Min_P^{KN} = \min_{VN.val \in P}\{VN.val: \exists_{SN \leftrightsquigarrow CN \leftrightsquigarrow KN} VN \leftrightsquigarrow SN\} \quad (3.25)$$

- Współczynnik reprezentatywności neuronu SN we wszystkich kombinacjach należących do AANG. Służy on do obniżenia wartości współczynnika dla dodanych neuronów SN.

$$\delta_{inSN}^{repr} = \left(\frac{1}{1+q_{inSN}^{repr}}\right)^2 \quad (3.26)$$

$$q_{inSN}^{repr} = \sum_{KN \in AANG} \|\{inSN: inSN \leftrightsquigarrow KN\}\| \quad (3.27)$$

- Współczynnik zawartych neuronów VN dla kandydującego neuronu SN, służy on do prezentacji liczby neuronów VN neuronu SN, które zostały już dodane do kombinacji.

$$\delta_{SN,KN}^{cont} = \left(\frac{1+q_{SN,KN}^{vn}}{\|\{RN:RN \leftrightsquigarrow KN\}\| + \|\{UN:UN \leftrightsquigarrow KN\}\|}\right)^2 \quad (3.28)$$

$$q_{SN,KN}^{vn} = q_{SN,KN}^{vnrn} + q_{SN,KN}^{vnun} \quad (3.29)$$

$$\delta_{SN,KN}^{vnrn} = \sum_{RN \leftrightsquigarrow KN} \|\{VN: VN \leftrightsquigarrow RN \wedge VN \leftrightsquigarrow SN \nleftrightarrow KN\}\| \quad (3.30)$$

$$\delta_{SN,KN}^{vnun} = \sum_{UN \leftrightsquigarrow KN} \|\{VN: VN \leftrightsquigarrow UN \wedge VN \leftrightsquigarrow SN \nleftrightarrow KN\}\| \quad (3.31)$$

W implementacji klasyfikatora ASONN do procesu tworzenia kombinacji dodano następujące kryteria w przypadku uzyskania współczynników o tej samej wartości.

- Minimalizacja współczynnika „Weeds" – w celu zminimalizowania negatywnego wpływu na poszerzanie kombinacji.
- Maksymalizacja współczynnika „Seeds" – w celu maksymalizacji pozytywnego wpływu na poszerzanie kombinacji.

3.2.2 Modyfikacje procesu tworzenia kombinacji

Modyfikacje procesu tworzenia kombinacji polegają na dodaniu dodatkowego pierwszego kryterium, które pozwala odrzucić część rozwiązań niespełniających zdefiniowanych warunków. W przypadku, gdy po zastosowaniu dodatkowego kryterium otrzymujemy więcej niż jedną możliwość, poszerzanie kombinacji kontynuowane jest według pierwotnego procesu tworzenia kombinacji. Na potrzeby pracy stworzono trzy alternatywne rozwiązania:

1. Minimalizacja obniżanego stopnia dyskryminacji:

Pierwsza modyfikacja polega na zapewnieniu, że kombinacja jest poszerzana w kierunku, który zapewnia obniżanie stopnia dyskryminacji jedynie w ostateczności. Wybierany jest współczynnik o najniższej wartości:

$$dirDiscrim_{P,KN}^{VN_i} = \frac{1}{\gamma_p} \cdot \left(\sum_{outSN \leftrightsquigarrow VN_i} snDiscrim_{KN}\right) \quad (3.32)$$

gdzie:

$$snDiscrim_{KN} = \|\{VN \leftrightsquigarrow SN: VN \leftrightsquigarrow RN \leftrightsquigarrow KN\} \cup \{VN \leftrightsquigarrow SN: VN \leftrightsquigarrow UN \leftrightsquigarrow KN\}\| \quad (3.33)$$

2. Maksymalizacja korelacji

Korelacje własne mają duży wpływ na poszerzanie kombinacji z tego względu, że wzorce posiadające tylko korelacje własne zawsze przynoszą korzyść dla reprezentatywności

kombinacji. Z tego powodu zdecydowano się na zastosowanie dwóch współczynników, z których wybierana jest maksymalna wartość. Pierwszy z nich bierze pod uwagę tylko korelacje własne, natomiast drugi bierze korelacja własne zredukowane o korelacje obce.

$$dirInCorr_{P,CN,KN}^{VN_i} = \frac{1}{\gamma_p} \cdot \left(\sum_{\substack{inSN \leftrightsquigarrow VN_i \wedge \\ inSN \leftrightsquigarrow CN \leftrightsquigarrow KN}} \|\{VN_i : VN_i \leftrightsquigarrow inSN\}\| \right) \tag{3.34}$$

$$dirReducedCorr_{P,CN,KN}^{VN_i} = \frac{1}{\gamma_p} \cdot \left(\sum_{\substack{inSN \leftrightsquigarrow VN_i \wedge \\ inSN \leftrightsquigarrow CN \leftrightsquigarrow KN}} \|\{VN_i : VN_i \leftrightsquigarrow inSN\}\| - \sum_{\substack{outSN \leftrightsquigarrow VN_i \wedge \\ outSN \nleftrightarrow CN \leftrightsquigarrow KN}} \|\{VN_i : VN_i \leftrightsquigarrow outSN\}\| \right) \tag{3.35}$$

3. Maksymalizacja SNVN

Trzeci sposób jest najbardziej wymagający obliczeniowo z tego względu, że dla wszystkich możliwych kierunków poszerzenia kombinacji i dla każdego inSN zawierającego VN sprawdzamy współczynniki z oryginalnej metody tworzenia kombinacji. Ma to na celu zdefiniowanie opłacalności dodawania danego kierunku na podstawie dalszych połączeń neuronu VN.

$$dirInSNVN_{P,CN,KN}^{VN_i} = \left(\sum_{\substack{inSN \leftrightsquigarrow VN_i \wedge \\ inSN \leftrightsquigarrow CN \leftrightsquigarrow KN}} \sum_{VN \leftrightsquigarrow inSN} \frac{1}{\gamma_p} * dir_{P,KN}^{VN} \right) \tag{3.36}$$

3.3 Budowa asocjacyjnego klasyfikatora ASONN

Ostatecznym etapem budowy asocjacyjnego klasyfikatora ASONN jest jego optymalizacja po utworzeniu kombinacji oraz dodaniu nowych neuronów KN, RN i UN oraz ich asocjacyjnych połączeń. Pierwszym etapem jest przekształcenie neuronów przedziału RN oraz neuronów podzbiorów UN w neurony receptoryczne, które będą pełnić funkcję wejścia reagującego na określone wartości parametrów. Neurony KN będą reagować na pobudzenia tych neuronów w zależności od wyznaczonej wagi połączenia odpowiednich neuronów z odpowiednimi neuronami KN. Następnie neurony CN będą reagować na pobudzenia neuronów KN. Natomiast wartość pobudzenia neuronu CN oznacza wynik klasyfikacji. Pobudzenia wyznaczane są na podstawie wartości wag, które obliczane są na podstawie poniższych wzorów.

Waga połączenia ADEF pomiędzy RN i KN:

$$w_{RN_k \leftrightsquigarrow KN} = \frac{\varphi_{RN_k}}{\sum_{RN_n \leftrightsquigarrow KN} \varphi_{RN_n} + \sum_{UN_m \leftrightsquigarrow KN} \varphi_{UN_m}} \tag{3.37}$$

$$\varphi_{RN_k} = \left(1 - \frac{Weeds_{RN_k}}{\varphi_{SN}^{discr}}\right) \cdot \frac{\varphi_{SN}^{repr} + Seeds_{RN_k}}{\varphi_{SN}^{repr} + AllSeeds_{KN}} \tag{3.38}$$

Waga połączenia ADEF pomiędzy UN i KN:

$$w_{UN_k \leftrightsquigarrow KN} = \frac{\varphi_{UN_k}}{\sum_{RN_n \leftrightsquigarrow KN} \varphi_{RN_n} + \sum_{UN_m \leftrightsquigarrow KN} \varphi_{UN_m}} \tag{3.39}$$

$$\varphi_{UN_k} = \left(1 - \frac{Weeds_{UN_k}}{\varphi_{SN}^{discr}}\right) \cdot \frac{\varphi_{SN}^{repr} + Seeds_{UN_k}}{\varphi_{SN}^{repr} + AllSeeds_{KN}} \tag{3.40}$$

gdzie:

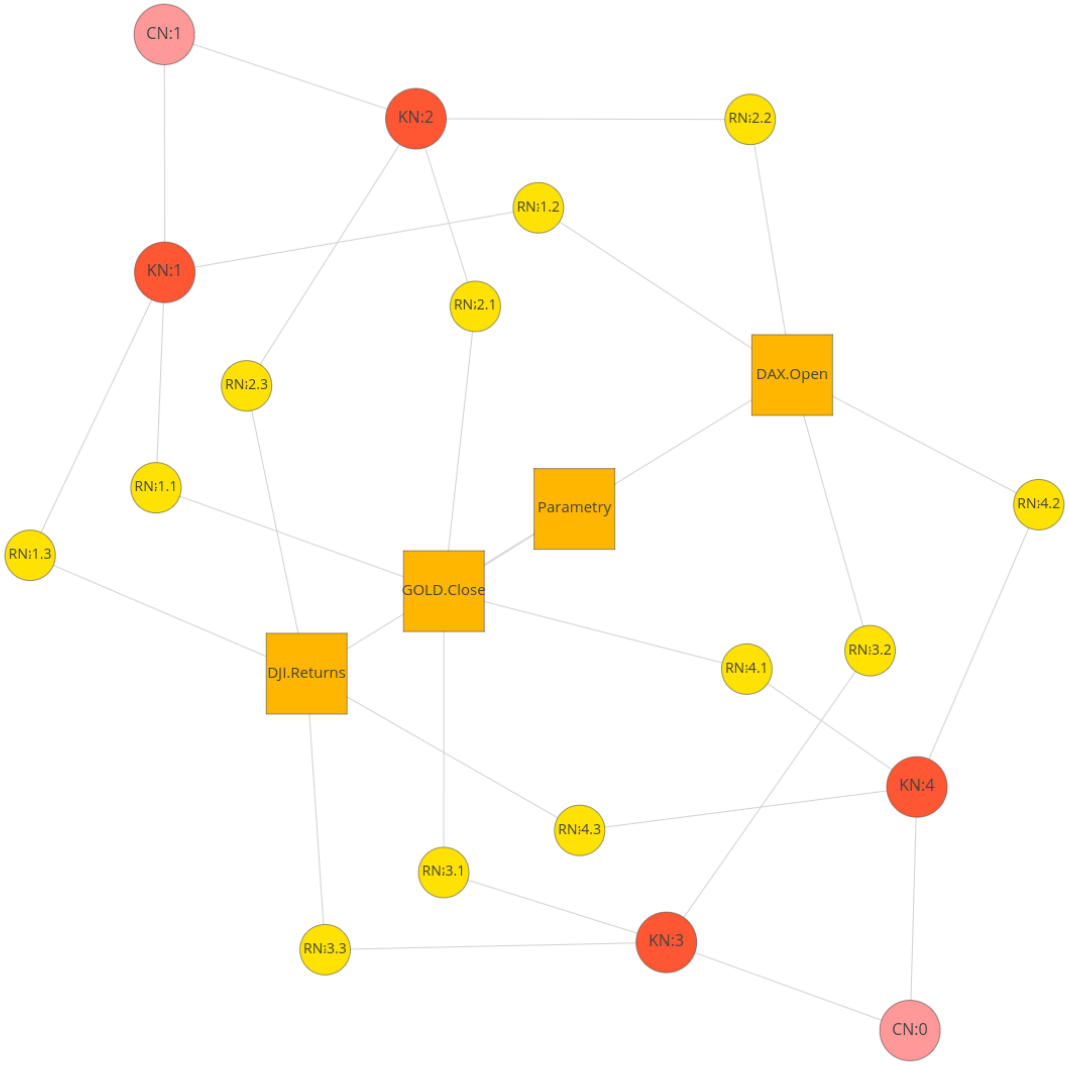
$$\varphi_{SN}^{repr} = \|\{inSN : inSN \leftrightsquigarrow KN\}\| \cdot \|\{RN : RN \leftrightsquigarrow KN\} \cup \{UN : UN \leftrightsquigarrow KN\}\|^2 \tag{3.41}$$

$$\varphi_{SN}^{discr} = \|\{SN : SN \nsim CN \leftrightsquigarrow KN\}\| \cdot \|\{RN : RN \leftrightsquigarrow KN\} \cup \{UN : UN \leftrightsquigarrow KN\}\|^2 \tag{3.42}$$

Dodatkowo neurony RN reagują na dane wejściowe według funkcji ściętego kapelusza Gaussa (3.43), co pozwala na pobudzenie neuronów przez wartości, które nie zawierają się w zakresie neuronu RN oraz pobudza z wartością 1 dla wartości znajdujących się w przedziale wartości.

$$TGH(x) = \begin{cases} 1, & RN_{Min} \leq x \leq RN_{Max} \\ e^{\frac{1-\left(\frac{2 \cdot x - RN_{Max} - RN_{Min}}{RN_{Max} - RN_{Min}}\right)^2}{2}}, & x < RN_{Min} \ \vee x > RN_{Max} \end{cases} \tag{3.43}$$

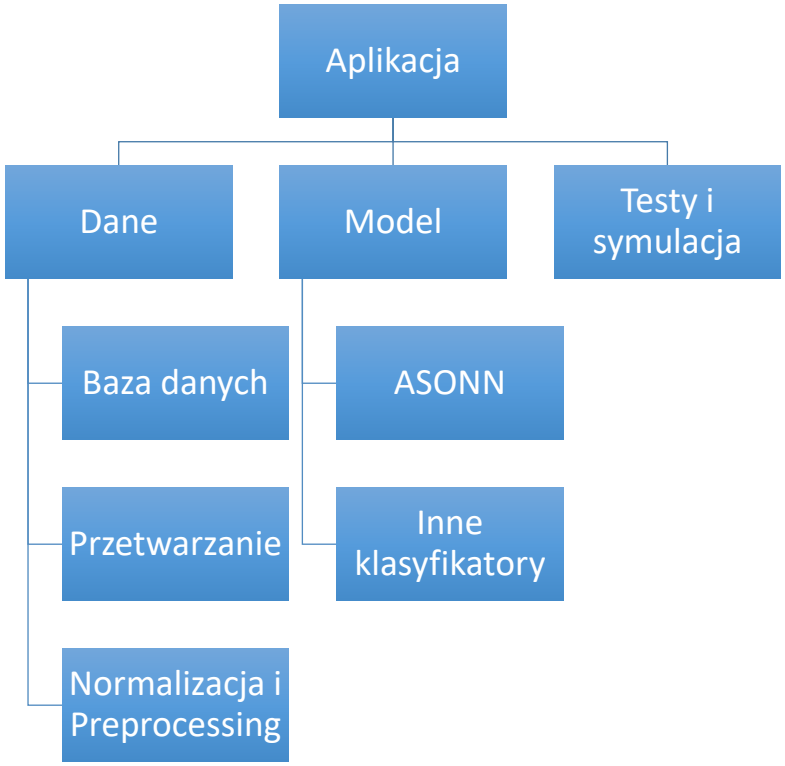
Po przeprowadzeniu całego procesu tworzenia kombinacji otrzymano strukturę klasyfikatora ASONN widoczną na rysunku 3.3.



Rysunek 3.3. Struktura klasyfikatora ASONN

4 Opis aplikacji

Stworzona w ramach pracy magisterskiej aplikacja została oparta na strukturze systemów wspomagania decyzji. Aplikacja została podzielona na trzy moduły, które zostały przedstawione na Rysunku 4.1 oraz opisane w kolejnych podrozdziałach.

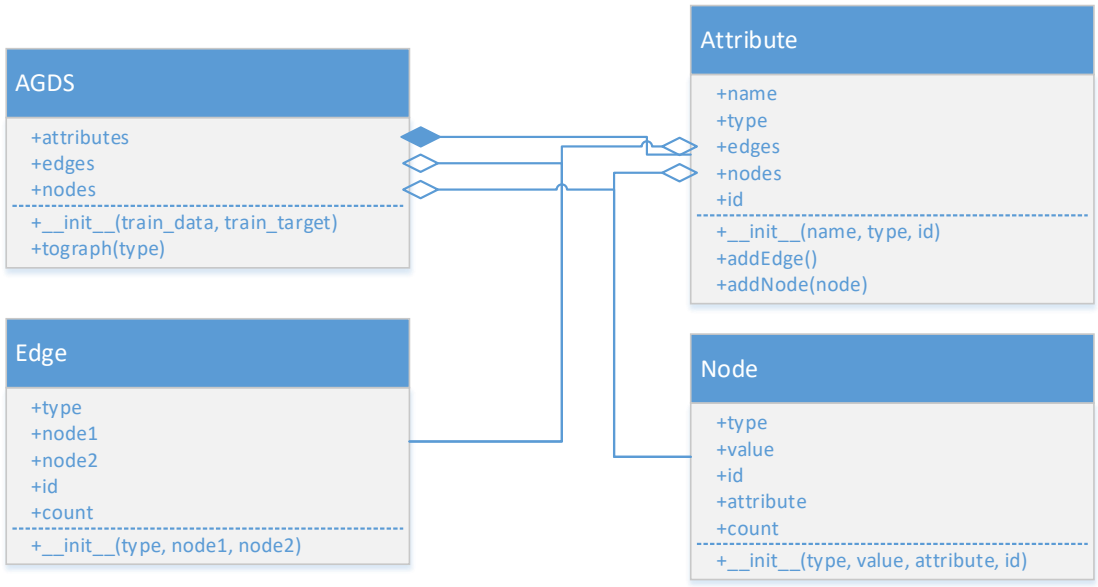


Rysunek 4.1. Struktura aplikacji

4.1 Model - implementacja ASONN

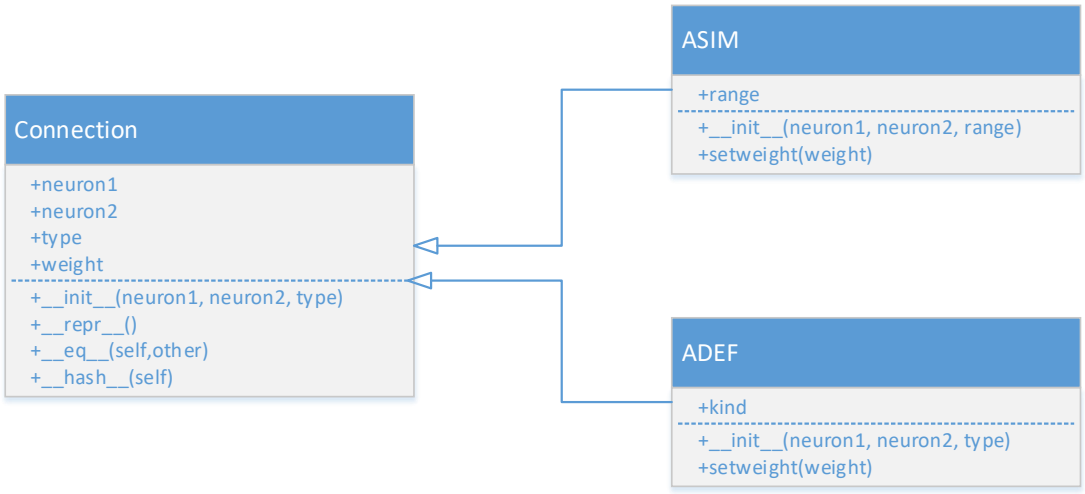
Nawiązując do poprzedniego rozdziału budowa klasyfikatora ASONN jest wieloetapowa i w ten sam sposób klasyfikator został zaimplementowany w aplikacji. Rozpoczynając od zbudowania AGDS, gdzie dane dostarczone w formie tabeli zostają przetworzone na strukturę grafową następująco: nazwy kolumn na klasę Attribute, która zawiera w sobie wszystkie wiersze, które reprezentowane są jako Node. Połączenia pomiędzy poszczególnymi Node są reprezentowane poprzez klasę Edge. Dzięki takiej prezentacji danych duplikaty oraz wzorce

sprzeczne są wykrywane w strukturze grafowej oraz odpowiednio przetwarzane w celu przekazania do AANG. Diagram klas został przedstawiony na rysunku 4.2

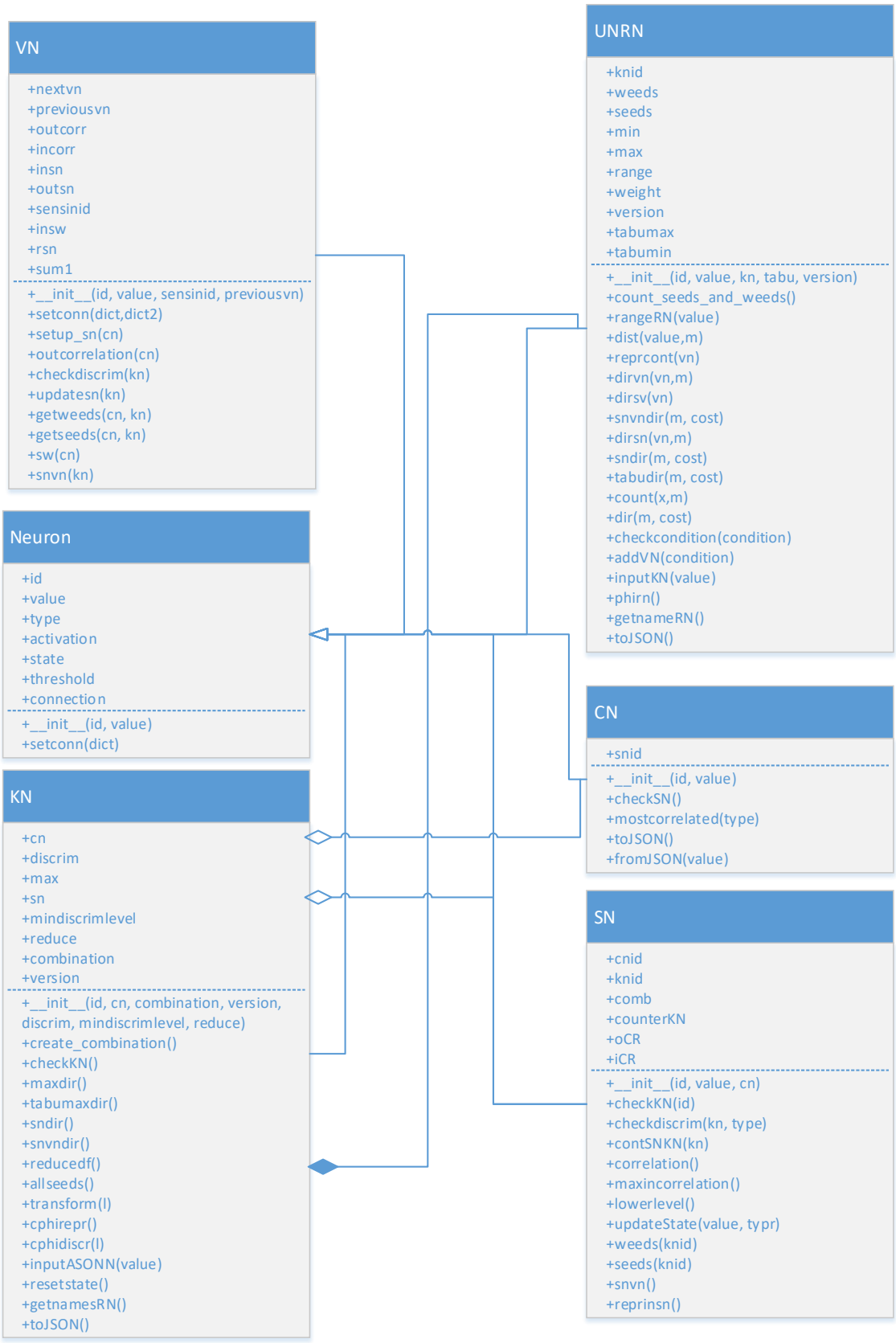


Rysunek 4.2. Diagram klas AGDS

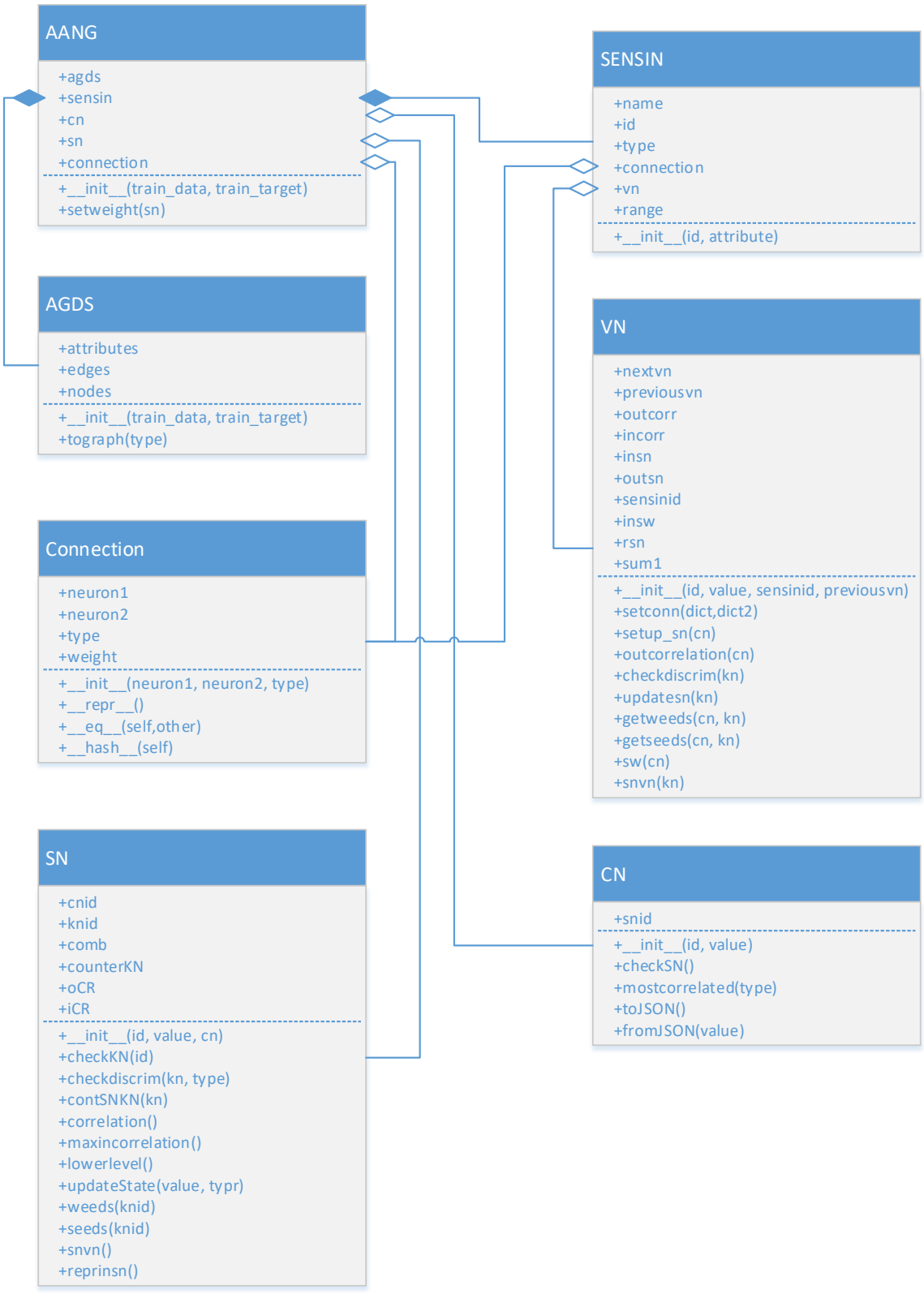
Kolejnym krokiem w budowie klasyfikatora ASONN jest zdefiniowanie podstawowych elementów grafu AANG opisanych w poprzednim rozdziale, czyli połączeń ASIM i ADEF (Rysunek 4.3), neuronów VN, SN, CN, KN i UNRN (Rysunek 4.4). Na podstawie tych elementów zdefiniowana SENSIN (wejście sensoryczne) oraz graf AANG (Rysunek 4.5).



Rysunek 4.3. Diagram klas Connection

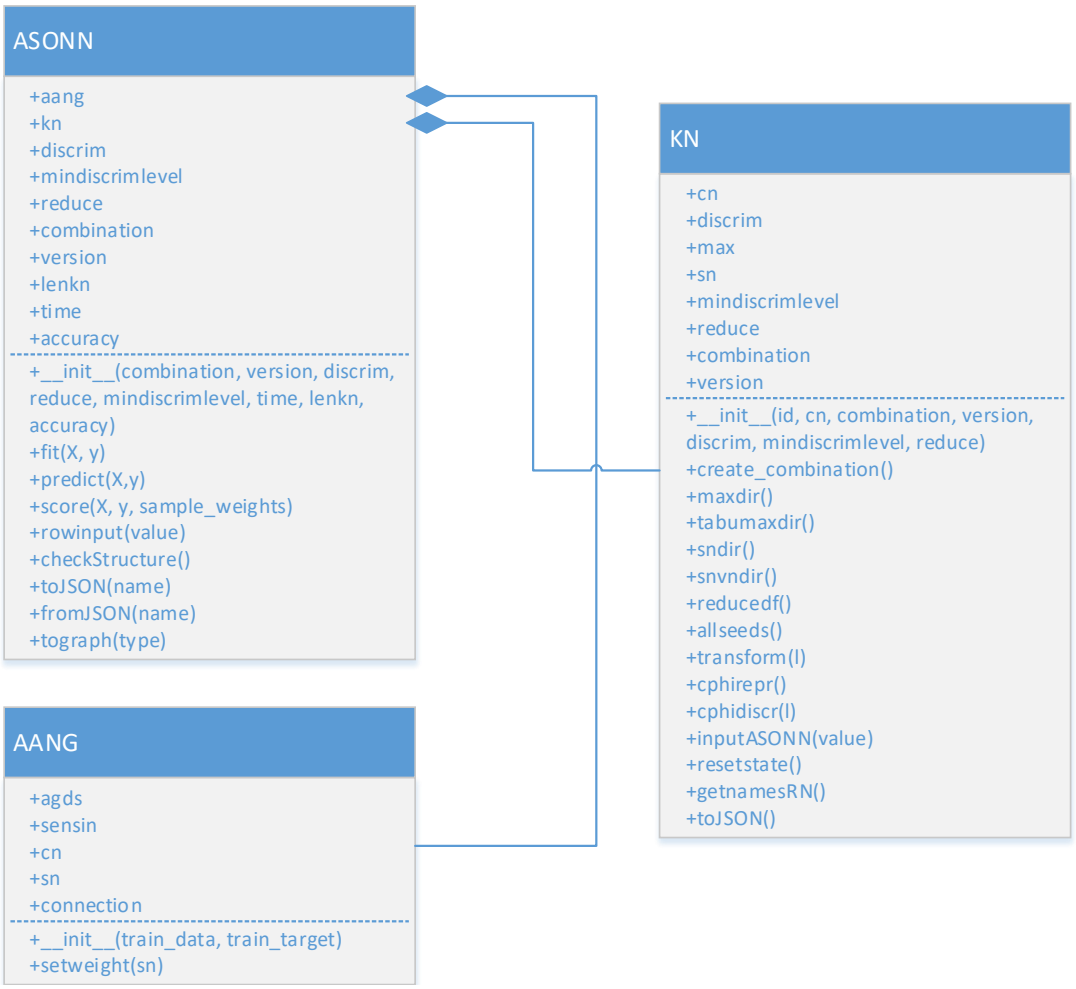


Rysunek 4.4. Diagram klas Neuron



Rysunek 4.5. Diagram klas AANG

Graf AANG zostaje stworzony na podstawie grafu AGDS. SENSIN zostaje stworzony na podstawie Attribute, Connection na podstawie Edge tworzy odpowiednie połączenia ASIM (połączenie pomiędzy dwoma neuronami VN) oraz ADEF (reszta połączeń). Natomiast neurony VN, SN oraz CN tworzone są na podstawie odpowiednich Node. W ten sposób uzyskany graf AANG umożliwia nam wyznaczenie wszystkich parametrów oraz zależności potrzebnych do utworzenia klasyfikatora ASONN.



Rysunek 4.6. Diagram klas ASONN

Końcowym krokiem jest tworzenie ASONN poprzez tworzenie kombinacji (neuron KN) na podstawie grafu AANG. Przy tworzeniu kombinacji bazujemy na wszystkich połączeniach oraz neuronach stworzonych w poprzednich krokach, przez co wykorzystujemy całą dostępną wiedzę z grafu AANG. Po utworzeniu kombinacji opisujących cały zbiór uczący, wyznaczane są wagi połączeń pomiędzy neuronami RN i KN, co pozwala nam uzyskać zoptymalizowaną i uproszczoną strukturę klasyfikatora ASONN. Klasyfikator został przystosowany do biblioteki Scikit-learn dla Pythona. Dzięki temu możemy korzystać z gotowych funkcjonalności

dotyczących klasyfikatorów między innymi poszukiwania optymalnych parametrów klasyfikatora.

4.2 Dane, testy oraz symulacja

W tym podrozdziale zostanie opisane podejście do uzyskania danych, następnie ich przechowywania oraz przetwarzania do postaci umożliwiającej wykorzystanie przez klasyfikatory.

Pierwszym krokiem jest uzyskanie danych oraz zapisanie ich w bazie danych. W tym celu wykorzystano serwis Quandl oraz dedykowane API dla Pythona, poprzez który zautomatyzowano proces pobierania danych. Dane niedostępne na Quandl pobierano za pomocą pliku csv z wielu serwisów między innymi World Bank, Eurostat i OECD. Baza danych została przystosowana do przechowywania danych w formacie OHLC dla szeregów czasowych oraz w formacie umożliwiającym przechowywanie danych fundamentalnych. Do stworzenia bazy danych wykorzystano PostgreSQL oraz do komunikacji SQLAlchemy (klasa Database Rysunek 4.7).

Kolejnym krokiem było przetwarzanie danych w celu uzyskania większej ilości informacji. Wszystkie wymienione w drugim rozdziale metody analizy technicznej oraz analizy statystycznej zostały zdefiniowane w klasie FinancialData, która jest odpowiedzialna za odpowiednie obliczenia, wyszukiwanie brakujących wartości oraz przygotowanie zbioru danych w postaci DataFrame w celu przygotowania zbioru danych do dalszego przetwarzania. Następnie w ten sposób przygotowany zbiór danych zostaje przekazany do klasy Data, która przystosowana jest do normalizacji, podziału zbiorów uczących i testowych oraz walidacji krzyżowej wykorzystywanej w bibliotece Scikit-learn. Dane przetworzone w klasie Data są przystosowane do użycia przez klasyfikatory. W celu uporządkowania rodzajów danych przekazywanych do klasyfikatorów wprowadzono klasę Template, w której definiowane są następujące parametry: nazwa, początek okresu czasowego, koniec okresu czasowego, zmienne (indeksy, waluty i inne) oraz ustawienia danych (parametry klasy Data dotyczące sposobu normalizacji oraz podziału na zbiór uczący i testowy). W ten sposób została zaimplementowana część aplikacji dotycząca przetwarzania danych wejściowych.

Przechodząc do testów konieczne było zaimplementowanie klasy Classifier, w której definiowany jest rodzaj klasyfikatora oraz jego parametry. Klasa ta udostępnia metody uczenia, testowania, walidacji krzyżowej oraz optymalizacji parametrów wejściowych stosowanych w bibliotece Scikit-learn. Jak wspomniano wcześniej klasyfikator ASONN także został dostosowany do tej biblioteki, co umożliwia łatwe porównanie wszystkich wybranych metod.

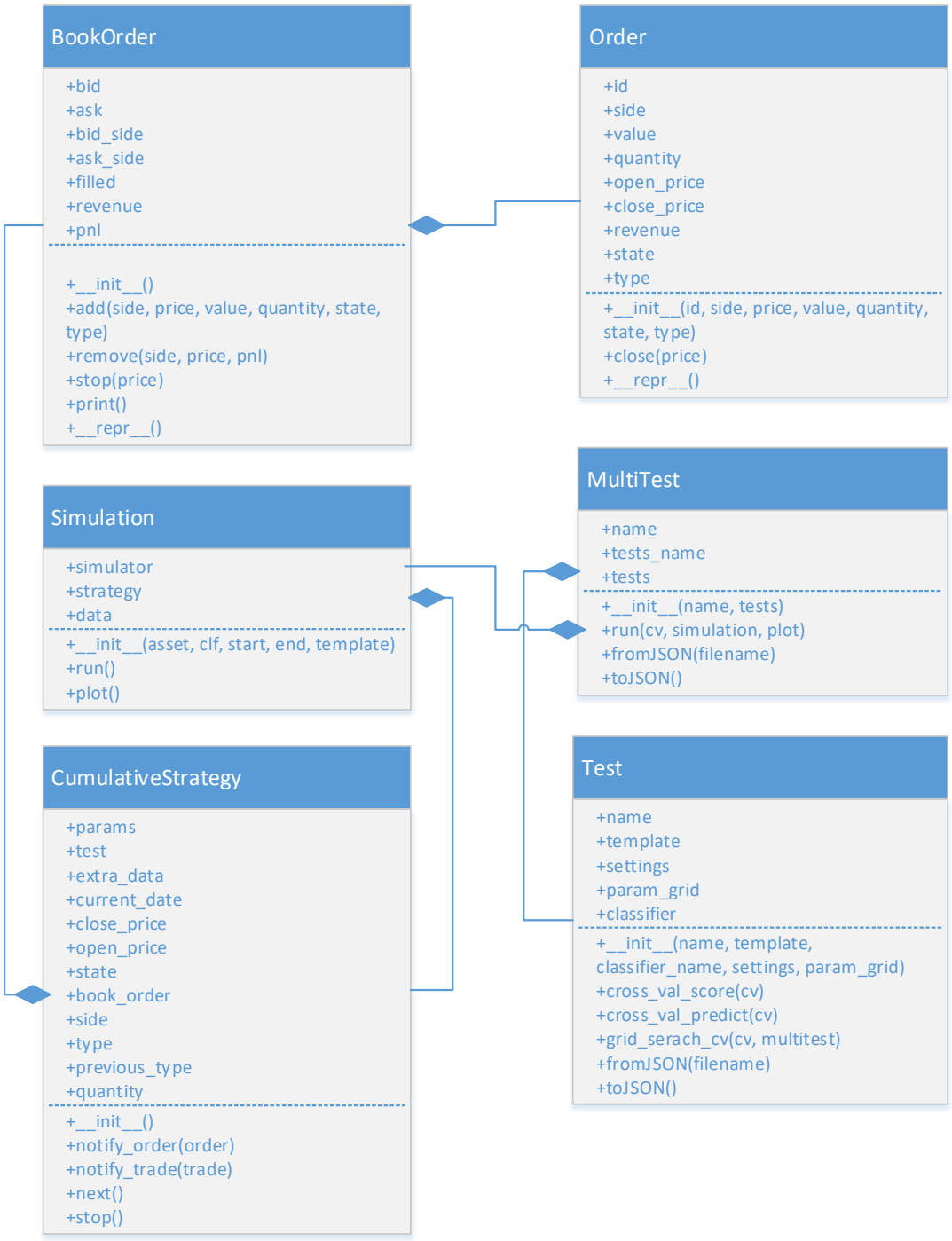
Rysunek 4.7. Diagram klas Data

Kolejnym krokiem w celu zautomatyzowania testów, było stworzenie klasy Test, w której definiujemy następujące parametry: nazwa, obiekt klasy Template, ustawienia klasyfikatora,

oraz zakres parametrów do optymalizacji. Dzięki tej klasie w jednym teście możemy zoptymalizować parametry jednego klasyfikatora. W celu porównania wszystkich klasyfikatorów oraz wygenerowania statystyk zdefiniowano klasę Multitest, która składa się z wcześniej opisanych testów. Klasa ta posiada metody umożliwiające zapis podsumowania działania klasyfikatorów do pliku JSON oraz wygenerowania danych statystycznych.

Jednakże w celu prezentacji możliwości klasyfikatorów na rynkach finansowych zdecydowano się na symulację działania klasyfikatorów na rynkach finansowych z wykorzystaniem biblioteki Backtrader. W tym celu stworzono klasę Simulation (Rysunek 4.8) zawierająca symulator z biblioteki Backtrader, dane potrzebne do przeprowadzenia symulacji oraz niezbędny element symulacji, którym jest strategia inwestycyjna (CumulativeStrategy), której zadaniem jest wykonywanie decyzji klasyfikatora. Rozpatrywano trzy możliwości decyzji: kupno, sprzedaż oraz zamknięcie pozycji w sytuacji braku pewności klasyfikacji. Na potrzeby przygotowania analizy transakcji stworzono klasy: BookOrder oraz Order. Klasy te umożliwiają określenie zysku/straty dla poszczególnej pozycji oraz całkowitego zysku/straty dla całej symulacji. Pozwalają także na prześledzenie stanu kapitału przez cały okres inwestycji, co pozwala na określenie ryzyka stosowania takiego klasyfikatora.

Implementacja aplikacji w ten sposób pozwoliła na zautomatyzowanie oraz zoptymalizowanie procesu testowania oraz symulacji, który ogranicza się do zdefiniowania danych wejściowych oraz wyboru odpowiednich metod klasyfikacji i doboru parametrów.



Rysunek 4.8. Diagram klas Simulation oraz Test

5 Testy

Celem pracy jest sprawdzenie przydatności klasyfikatora ASONN w procesie wspomagania decyzji inwestycyjnych oraz osiągnięcie jak najlepszych wyników klasyfikacji w porównaniu z innymi metodami uczenia maszynowego. Testy mają na celu pokazanie, jaki rodzaj danych i o jakim rozkładzie najlepiej wpływają na proces tworzenia kombinacji, która wpływa na skuteczność klasyfikacji oraz reprezentatywność kombinacji. W tym celu przeprowadzono szerokie testy stosując następującą metodologię:

1. Dla każdej metody stosowana jest odpowiednia normalizacja danych, 10-krotna walidacja krzyżowa oraz wyznaczane są dane statystyczne, na których podstawie metody są porównywane. Obliczane są następujące dane statystyczne: skuteczność klasyfikacji, zwrot z inwestycji, ilość transakcji zyskownych (Z) oraz stratnych (S) oraz stosunek transakcji zyskownych do ilości wszystkich transakcji.
2. Dla każdej metody wyznaczany jest optymalny zestaw parametrów za pomocą klasy Test i metody Grid_Search.
3. Testy przeprowadzane są na różnych okresach czasowych od 2015 roku.
4. Symulacje działania klasyfikatorów przedstawiane są na danych z 2017 roku.
5. Zbiór danych wejściowych jest wyznaczany stopniowo zaczynając od testów zawierających jeden rodzaj danych tj. dane z analizy technicznej, dane fundamentalne oraz dane statystyczne. Na podstawie uzyskanych wyników stworzono optymalną strukturę danych wejściowych.
6. Porównanie dla wszystkich metod zostały zautomatyzowane poprzez użycie klasy Multitest.

Poprawność działania klasyfikatora ASONN została przetestowana na zbiorze IRIS, gdzie osiągnięto 99% skuteczności oraz na zbiorze Wine, gdzie osiągnięto 98% skuteczności w obu przypadkach przy 10-krotnej walidacji krzyżowej. W tabelach opisujących wyniki przyjęto następujące oznaczenia:

- ASONN-VN jako oryginalna metoda tworzenia kombinacji,
- ASONN-DISCRIM jako minimalizacja stopnia dyskryminacji,
- ASONN-SNVN jako maksymalizacja SNVN,
- ASONN-CORR jako maksymalizacja korelacji,
- DT jako drzewo decyzyjne,
- KNN jako k najbliższych sąsiadów,
- NN jako sieci neuronowe,
- SVM jako maszyna wektorów nośnych.

5.1 S&P 500

W tym podrozdziale przedstawiono testy klasyfikatorów dla indeksu S&P 500. Zgodnie z opisywaną metodologią poszukiwania optymalnego zbioru danych wejściowych rozpoczęto od poszukiwania wskaźników analizy technicznej. Na podstawie wyników różnego rodzaju testów wybrano następujące wskaźniki obliczone na podstawie cen S&P500: ADX, ATR, EMA, CCI, RSI, STOCH oraz OBV. Uzyskane wyniki przedstawiono w tabeli 5.1. W celu porównania metod uczenia maszynowego, największą uwagę zwrócono na skuteczność klasyfikacji podczas walidacji krzyżowej, stopę zwrotu oraz stosunek transakcji zyskownych do wszystkich przeprowadzonych transakcji, które pokazują zdolności klasyfikacji wybranych klasyfikatorów (im większe wartości tym lepsza ocena klasyfikatora). W pierwszym teście najlepsze wyniki uzyskano dla NN oraz SVM, dla których osiągnięto najwyższą stopę zwrotu o wartości 6,7% oraz stosunek zyskownych transakcji o wartości 0,8. Analizując dane z tabeli można dojść do wniosku, że skuteczność klasyfikacji uzyskana podczas walidacji krzyżowej nie odzwierciedla wyniku finansowego uzyskanego przez daną metodę uczenia maszynowego podczas symulacji rynku finansowego, na co wskazuje porównanie kolumn „Skuteczność" oraz „Stopa zwrotu" (0,56 dla NN i stopa zwrotu 6,7% w porównaniu do ASONN-CORR odpowiednio 0,6 i 0,2%). Poniższe dane wskazują również, że w przypadku, gdy algorytm podejmuję więcej stratnych decyzji, możliwe jest uzyskanie zysku na koniec inwestycji (ASONN-CORR stopa zwrotu 0,2% dla 48% zyskownych transakcji). Przechodząc do klasyfikatora ASONN, stopa zwrotu była ujemna tylko podczas tworzenia kombinacji typu VN, jednakże stopy zwrotu dla pozostałych typów również znacznie odbiegają od dwóch najlepszych wyników uzyskanych w tym teście.

Tabela 5.1. Wyniki klasyfikacji dla danych z analizy technicznej

Klasyfikator	**Skuteczność**	**Stopa zwrotu**	**Ilość transakcji (Z)**	**Ilość transakcji (S)**	**Stosunek Z/Z+S**
ASONN-VN	0,51	-0,01%	50	44	0,53
ASONN-DISCRIM	0,51	0,92%	54	40	0,57
ASONN-SNVN	0,53	2,1%	59	35	0,63
ASONN-CORR	0,6	0,2%	38	41	0,48
DT	0,51	-0,7%	32	62	0,34
KNN	0,54	0,19%	57	37	0,6
NN	0,56	6,7%	75	19	0,8
SVM	0,57	6,7%	75	19	0,8

Kolejnym etapem poszukiwania są testy z użyciem danych fundamentalnych. Testy te podzielono na trzy grupy: stopy procentowe, wskaźniki koniunktury oraz indeksy pewności i wyceny. Do każdej grupy zostały dodane także dane surowcowe. W pierwszej grupie klasyfikator ASONN uzyskał znaczną przewagę nad pozostałymi metodami przede wszystkim pod względem stopy zwrotu oraz stosunku transakcji zyskownych. Najlepszy wynik uzyskano dla ASONN-VN, który uzyskał skuteczność 0,63, stopę zwrotu na poziomie 2,3% oraz przeprowadził 70% zyskownych transakcji. Dane fundamentalne mają zdecydowanie inny rozkład wartości w stosunku do danych uzyskanych z metod analizy technicznej. Klasyfikator ASONN daje lepsze wyniki dla nie sprzecznych danych uczących, które w dużym stopniu obecne są na rynkach finansowych. Z tej grupy wybrano stopę procentową Euribor, 5 letnie Treasury Bond oraz cenę złota.

Tabela 5.2. Wyniki klasyfikacji dla grupy stopy procentowe

Klasyfikator	**Skuteczność**	**Stopa zwrotu**	**Ilość transakcji (Z)**	**Ilość transakcji (S)**	**Stosunek Z/Z+S**
ASONN-VN	0,63	2,3%	42	18	0,7
ASONN-DISCRIM	0,54	1%	41	19	0,68
ASONN-SNVN	0,56	2,2%	41	19	0,68
ASONN-CORR	0,53	1,2%	34	26	0,56
DT	0,54	-0,1%	30	30	0,5
KNN	0,59	-0,2%	22	38	0,37
NN	0,58	-0,5%	22	38	0,37
SVM	0,57	-0,5%	20	40	0,33

Tabela 5.3. Wyniki klasyfikacji dla grupy wskaźniki koniunktury

Klasyfikator	**Skuteczność**	**Stopa zwrotu**	**Ilość transakcji (Z)**	**Ilość transakcji (S)**	**Stosunek Z/Z+S**
ASONN-VN	0,54	-0,3%	12	27	0,31
ASONN-DISCRIM	0,57	2,1%	34	5	0,87
ASONN-SNVN	0,6	2,3%	36	3	0,92
ASONN-CORR	0,56	2,3%	35	4	0,9
DT	0,59	0,04%	17	22	0,44
KNN	0,6	0,2%	25	14	0,64
NN	0,57	0,1%	27	12	0,69
SVM	0,51	2,2%	35	4	0,9

Przechodząc do drugiej grupy, w porównaniu do pierwszej pozostałe metody uzyskały znacznie lepsze wyniki, jednakże pod względem stopy zwrotu klasyfikator ASONN znacznie przeważał (oprócz oryginalnej metody VN). W tej grupie największy wpływ na klasyfikacje miała: inflacja, stopa bezrobocia oraz produkcji przemysłowa. Są to wskaźniki, które podczas publikacji także mają bardzo wysoki wpływ na rynki finansowe, dlatego też poniższe wyniki potwierdzają tą zależność. W ostatniej grupie większość metod przynosi stratę oraz nie ma pozytywnego wpływu na inwestycję, dlatego też zrezygnowano z tej grupy danych w dalszej optymalizacji struktury danych.

Tabela 5.4. Wyniki klasyfikacji dla grupy indeksy pewności i wyceny

Klasyfikator	**Skuteczność**	**Stopa zwrotu**	**Ilość transakcji (Z)**	**Ilość transakcji (S)**	**Stosunek Z/Z+S**
ASONN-VN	0,53	-0,01%	11	20	0,35
ASONN-DISCRIM	0,58	0,5%	21	19	0,525
ASONN-SNVN	0,65	0,5%	21	19	0,525
ASONN-CORR	0,67	-0,3%	20	20	0,5
DT	0,55	-0,2%	15	25	0,375
KNN	0,55	-0,6%	16	24	0,4
NN	0,5	0,001%	23	17	0,58
SVM	0,51	-0,5%	19	21	0,475

Następnie użyto danych statystycznych. Dla różnych metod tworzenia klasyfikatora ASONN uzyskano znacznie różniące się wyniki. Z tabeli 5.5 wynika, że klasyfikator może mieć problem z danymi numerycznymi, które w zdecydowanej większości nie są powtarzalne w zbiorze danych. Z danych statystycznych zdecydowano się na wybranie współczynnika beta oraz standardowe odchylenie kwadratowe stopy zwrotów.

Podsumowując, na podstawie przeprowadzonych testów zdefiniowano następujący optymalny zbiór danych wejściowych:

1. Wskaźniki analizy technicznej: ADX, ATR, EMA, CCI, RSI, STOCH oraz OBV.
2. Dane fundamentalne: inflacja, stopa bezrobocia, produkcji przemysłowa, Euribor, 5-letnie Treasury Bond oraz złoto.
3. Dane statystyczne: współczynnik beta oraz standardowe odchylenie kwadratowe.

Tabela 5.5. Wyniki klasyfikacji dla danych statystycznych.

Klasyfikator	**Skuteczność**	**Stopa zwrotu**	**Ilość transakcji (Z)**	**Ilość transakcji (S)**	**Stosunek Z/Z+S**
ASONN-VN	0,67	2,3%	45	13	0,77
ASONN-DISCRIM	0,65	0,8%	28	30	0,48
ASONN-SNVN	0,5	-0,8%	30	28	0,52
ASONN-CORR	0,5	-2%	15	43	0,26
DT	0,55	-0,4%	23	35	0,4
KNN	0,53	0,3%	35	23	0,6
NN	0,52	1,5%	35	23	0,6
SVM	0,6	2%	39	19	0,67

Wyniki uzyskane za pomocą wyznaczonej optymalnej struktury danych przedstawiono w tabeli 5.6.

Tabela 5.6. Wyniki klasyfikacji optymalnej struktury danych

Klasyfikator	**Skuteczność**	**Stopa zwrotu**	**Ilość transakcji (Z)**	**Ilość transakcji (S)**	**Stosunek Z/Z+S**
ASONN-VN	0,65	5,1%	70	21	0,77
ASONN-DISCRIM	0,62	3,8%	64	30	0,68
ASONN-SNVN	0,6	4,8%	65	25	0,72
ASONN-CORR	0,61	3,1%	58	32	0,64
DT	0,53	-0,2%	37	51	0,42
KNN	0,51	0,1%	40	50	0,44
NN	0,56	2,9%	55	37	0,60
SVM	0,62	4%	64	28	0,70
Średnia	**0,59**	**2,95%**	**57**	**34**	**0,62**

Z powyższej tabeli wynika, że dzięki odpowiedniemu zbiorowi danych wejściowych uzyskano znacznie lepsze wyniki w przypadku klasyfikatora ASONN. Wybrane dane pokazują, że najlepsze efekty przynoszą dane, które na podstawie wartości określają zdefiniowany stan lub sygnał, np. wskaźnik RSI, który dla poziomu niższego niż 30 oznacza sygnał kupna, a na poziomie wyższym niż 70 sygnał sprzedaży. Testy dla indeksu giełdowego S&P 500 pokazały, że większość metod (oprócz DT) może być zyskowna na rynkach finansowych. Średnia uzyskana stopa zwrotu dla wszystkich metod wynosi 2,95%, natomiast dla klasyfikatora ASONN średnia stopa zwrotu wynosi 4,2%. Średnia uzyskana skuteczność wynosi 0,59, natomiast stosunek transakcji zyskownych średnio wynosi 0,62.

5.2 Kurs pary walutowej EUR/USD

W tym podrozdziale przedstawiono testy klasyfikatorów dla pary walutowej EUR/USD. Etapy testów wyglądają tak samo jak dla indeksu S&P 500. W pierwszym teście użyto różne kombinacji danych z analizy technicznej. Najlepsze wyniki uzyskano dla następujących danych obliczanych za pomocą ceny pary walutowej EUR/USD: WMA, CCI, MACD, MFI, RSI, ATR oraz OBV. Porównanie wyników dla klasyfikatorów przedstawiono w tabeli 5.7. Wszystkie metody uczenia maszynowego uzyskały pozytywną stopę zwrotu o zdecydowanie wyższych wartościach w porównaniu do S&P 500 (nawet 7,4%). Dlatego też, bardzo istotnym elementem podczas tworzenia automatycznego systemu tradingowego jest dobór odpowiedniej metody do rynku finansowego.

Tabela 5.7. Wyniki klasyfikacji dla danych z analizy technicznej

Klasyfikator	**Skuteczność**	**Stopa zwrotu**	**Ilość transakcji (Z)**	**Ilość transakcji (S)**	**Stosunek Z/Z+S**
ASONN-VN	0,55	7,4%	59	40	0,6
ASONN-DISCRIM	0,63	8%	64	35	0,65
ASONN-SNVN	0,54	7,4%	59	40	0,6
ASONN-CORR	0,54	7,4%	59	40	0,6
DT	0,63	1,5%	37	62	0,37
KNN	0,55	4%	50	49	0,51
NN	0,65	6%	57	42	0,58
SVM	0,65	3,4%	45	54	0,45

Następnie przeprowadzono testy z użyciem danych fundamentalnych. Testy te zostały podzielone w ten sam sposób jak w przypadku indeksu S&P500, a następnie przeprowadzono test klasyfikatora składającego się z wszystkich wybranych danych fundamentalnych. Ze względu na dużą zależność pomiędzy indeksem giełdowym S&P500, a kursem walutowym EURUSD nie dziwi fakt, że największy wpływ miały te same dane fundamentalne, które zostały wyznaczone w poprzednim podrozdziale. Jednakże wpływ tych zmiennych na pozytywny wynik jest dużo mniejszy dla klasyfikatora ASONN w porównaniu do użycia danych z analizy technicznej, a dla NN i SVM przynosi negatywny efekt. Wyniki przedstawiono w poniższej tabeli.

Tabela 5.8. Wyniki klasyfikacji dla danych fundamentalnych.

Klasyfikator	**Skuteczność**	**Stopa zwrotu**	**Ilość transakcji (Z)**	**Ilość transakcji (S)**	**Stosunek Z/Z+S**
ASONN-VN	0,55	2,4%	55	44	0,56
ASONN-DISCRIM	0,63	1,8%	50	49	0,51
ASONN-SNVN	0,54	1,9%	51	48	0,52
ASONN-CORR	0,54	1,5%	50	49	0,51
DT	0,63	0,3%	42	57	0,42
KNN	0,55	0,1%	44	55	0,44
NN	0,65	-0,1%	39	60	0,39
SVM	0,65	-0,4%	35	64	0,35

Przechodząc do danych statystycznych, podobnie jak dla indeksu S&P500 dla różnych metod tworzenia klasyfikatora ASONN uzyskano znacznie różniące się wyniki. Największy wpływ na wynik miał współczynnika beta oraz standardowe odchylenie kwadratowe stóp zwrotu, dlatego też zdecydowano się na jego wykorzystanie w końcowej strukturze danych wejściowych.

Tabela 5.9. Wyniki klasyfikacji dla danych statystycznych.

Klasyfikator	**Skuteczność**	**Stopa zwrotu**	**Ilość transakcji (Z)**	**Ilość transakcji (S)**	**Stosunek Z/Z+S**
ASONN-VN	0,57	-1,1%	20	40	0,33
ASONN-DISCRIM	0,56	0,2%	12	20	0,38
ASONN-SNVN	0,55	1,3%	27	33	0,45
ASONN-CORR	0,57	-1,1%	20	40	0,33
DT	0,59	0,5%	29	31	0,48
KNN	0,56	0,5%	26	34	0,43
NN	0,54	0,5%	27	33	0,45
SVM	0,53	1,3%	27	33	0,45

Podsumowując, na podstawie przeprowadzonych testów zdefiniowano następujący optymalny zbiór danych wejściowych:

1. Wskaźniki analizy technicznej: WMA, CCI, MACD, MFI, RSI, ATR oraz OBV.

2. Dane fundamentalne: inflacja, stopa bezrobocia, produkcji przemysłowa, Euribor, 5-letnie Treasury Bond oraz złoto.
3. Dane statystyczne: współczynnik beta oraz standardowe odchylenie kwadratowe stopy zwrotu.

Wyniki uzyskane za pomocą wyznaczonej optymalnej struktury danych dla EUR/USD przedstawiono w tabeli 5.10.

Tabela 5.10. Wyniki klasyfikacji optymalnej struktury danych

Klasyfikator	**Skuteczność**	**Stopa zwrotu**	**Ilość transakcji (Z)**	**Ilość transakcji (S)**	**Stosunek Z/Z+S**
ASONN-VN	0,63	7,5%	67	35	0,66
ASONN-DISCRIM	0,64	8%	64	35	0,65
ASONN-SNVN	0,61	7,4%	62	40	0,61
ASONN-CORR	0,59	7,6%	57	38	0,60
DT	0,64	1,2%	40	62	0,39
KNN	0,52	3%	50	50	0,50
NN	0,66	6,5%	57	42	0,58
SVM	0,67	5,4%	66	39	0,63
Średnia	0,62	5,8%	58	43	0,58

Porównując do wcześniej uzyskanych wyników z analizy technicznej nie ma dużej różnicy zarówno w stopie zwrotu, skuteczności jak i stosunku transakcji zyskownych i stratnych. Pokazuje to, że dla kursu walutowego EUR/USD duże znaczenie mają wskaźniki analizy technicznej, które pokazują, w którym miejscu cyklu znajduje się cena. Natomiast bardzo niskie znaczenie mają dane fundamentalne oraz statystyczne. Może to wynikać z faktu, że dane fundamentalne mają zbyt niską zmienność (dane miesięczne, a rozpatrywane są inwestycje o horyzoncie czasowym dziennym). Wyniki klasyfikacji uzyskane dla klasyfikatora ASONN okazały się dużo lepsze od innych metod uczenia maszynowego (stopa zwrotu minimum 7,4%), mimo tego, że nie uzyskały najlepszej skuteczności podczas walidacji krzyżowej. Najlepszy wyniki uzyskano dla ASONN-Discrim, natomiast najgorszy dla DT (1,2% mimo wysokiej skuteczności).

5.3 Podsumowanie testów

W celu zbadania skuteczności klasyfikatora ASONN oraz sposobów tworzenia kombinacji przeprowadzono testy sprawdzające użyteczność klasyfikatora pod wieloma względami. Pierwszym z nich było sprawdzenie, jaki rodzaj danych dostarczony do klasyfikatora umożliwi uzyskanie wysokich skuteczności podczas procesu walidacji krzyżowej. Następnie testowano czas uczenia w zależności od wielkości zbioru danych wejściowych oraz liczby danych wejściowych. Końcowym etapem badań była symulacja działania klasyfikatorów w warunkach rynkowych. Na podstawie tych elementów wyciągnięto następujące wnioski:

1. Skuteczność podczas walidacji krzyżowej w wielu przypadkach nie przesądza o skuteczności działania podczas symulacji.
2. Wszystkie testowane metody posiadały pozytywne stopy zwrotu w zależności od rodzaju użytych danych (bez uwzględnienia prowizji, co przy dużej częstotliwości transakcji może znacznie obniżyć stopę zwrotu).
3. Klasyfikator ASONN przy odpowiednich danych wejściowych przynosi bardzo dobre efekty przy wspomaganiu podejmowania decyzji inwestycyjnych.
4. Największą wadą klasyfikatora ASONN jest czas uczenia dla dużych nisko skorelowanych zbiorów danych, dlatego też podstawowym elementem dla wyznaczenia optymalnego klasyfikatora jest dobór odpowiednich danych.

6 Podsumowanie

Podstawowym celem pracy była implementacji oraz sprawdzenie przydatności klasyfikatora ASONN w systemie wspomagania decyzji inwestycyjnych na rynkach finansowych, które są niezwykle wymagające dla inwestorów. W tym celu stworzono system wspomagania decyzji, który składa się z bazy danych zawierającej dane finansowe oraz modeli, którymi są klasyfikator ASONN oraz inne wykorzystane metody uczenia maszynowego. Kolejnymi elementami systemu są: środowisko pozwalające na przetwarzaniu danych finansowych, środowisko testowe pozwalające na optymalizację klasyfikatorów oraz środowisko symulacyjne pozwalające na sprawdzenie wyniku finansowego przy użytkowaniu danego klasyfikatora.

Korzystając z wyżej wymienionych funkcjonalności przeprowadzono testy oraz symulacje, z których wyciągnięto następujące wnioski. Pierwszym z nich jest, że metody uczenia maszynowego z powodzeniem można stosować przy wspomaganiu decyzji na rynkach finansowych, przy czym niezwykle istotny jest odpowiedni dobór danych oraz ich formatu w zależności od stosowanej metody. Przechodząc do klasyfikatora ASONN, każda z metod tworzenia kombinacji zaimplementowanych na potrzeby tej pracy ma swoje wady i zalety i ich skuteczność zależy od dostarczonych danych wejściowych. Największą wadą jest długi czas tworzenia kombinacji dla dużych zbiorów danych, jednakże w tym przypadku jest dużo możliwości optymalizacji w porównaniu do zastosowanej w tej pracy implementacji. Przede wszystkim większość obliczeń można wykonać równolegle, co może znacznie przyspieszyć proces tworzenie kombinacji. Drugim sposobem jest optymalizacja kodu oraz wykorzystanie bibliotek do obliczeń numerycznych, co może znacznie przyspieszyć krytyczne fragmentu kodu ze punktu widzenia tworzenia kombinacji. Kolejnym bardzo istotnym czynnikiem jest podejście do uczenia i walidacji, ponieważ jak wykazano w testach skuteczność uzyskana podczas walidacji krzyżowej nie koniecznie pokrywa się ze skutecznością klasyfikatora podczas symulacji, co odwzorowuje wykorzystanie klasyfikatora na rzeczywistym rynku. W takich sytuacjach decyzja o wybraniu danego klasyfikatora jest znacznie utrudniona, przez co konieczne wydaje się wybranie alternatywnych metod oceny skuteczności klasyfikatorów. Z pomocą w tym zakresie mogą przyjść metody zarządzania ryzykiem oraz zarządzeniem portfelem inwestycyjnym. Podejście to wymaga stworzenia automatycznego systemu tradingowego, który zawiera w sobie nie tylko model generujący decyzje inwestycyjne, ale również system zarządzania pozycją inwestycyjną, zarządzania ryzykiem oraz przetwarzaniem danych w czasie rzeczywistym.

Podsumowując, podejście do tworzenia klasyfikatora ASONN poprzez różne sposoby tworzenia kombinacji przyniosło pozytywny skutek w postaci pozytywnej stopy zwrotu podczas symulacji na rynkach finansowych. W porównaniu do innych metod uczenia maszynowego uzyskiwał skuteczność na podobnym lub wyższym poziomie. Na podstawie

obserwacji skuteczności podczas testowania klasyfikatora ASONN zauważono wysoką skuteczność dla danych numerycznych zawierających się w ograniczonym przedziale, których wartość ma określone znaczenie podczas procesu podejmowania decyzji. Prowadzi to do wniosków, że klasyfikator ASONN może mieć wysoką skuteczność przy agregacji innych prognoz dotyczących decyzji inwestycyjnych między innymi z wykorzystaniem metod uczenia maszynowego, analizy statystycznej lub analizy fundamentalnej. Pokazuje to niezwykle intersujący kierunek dalszych badań rozwijających klasyfikator ASONN.

Przechodząc do porównania uzyskanych wyników z pracami naukowymi, zaobserwowano znaczne różnice dla drzew decyzyjnych oraz k najbliższych sąsiadów, które miały dużo mniejszą skuteczność. Powodem tego może być ukierunkowanie na optymalizację klasyfikatora ASONN, którego skuteczność jest na poziomie porównywalnym do najlepszych metod uczenia maszynowego wykorzystywanego na rynkach finansowych.

7 Bibliografia

[1] B. Krollner, B. Vanstone i G. Finnie, „Financial time series forecasting with machine learning techniques: A survery," w *European Symposium on Artificial Neural Networks: Computational and Machine Learning*, Bruges, Belgium, 2010.

[2] A. Horzyk, Sztuczne systemy skojarzeniowe i asocjacyjna sztuczna inteligencja, Warszawa: Akademicka Oficyna Wydawnicza EXIT, 2013.

[3] B. W. Weber, „Financial DSS: Systems for supporting investment decisions.," *Handbook on Decision Support Systems 2,* pp. 419-442, 2008.

[4] D. Dymek, W. Komnata, L. Kotulski i P. Szwed, Architektury Hurtowni Danych Model referencyjny i formalny opis architektury, Kraków: Wydawnictwa AGH, 2015.

[5] J. J. Murphy, Międzyrynkowa Analiza Techniczna Strategie inwestycyjne na rynkach Akcji, Obligacji, Towarów i Walut, Warszawa: WIG-Press, 1998.

[6] *Ustawa z dnia 29 września 1994 r. o rachunkowości. Dz. U. 2016, poz. 1047, art. 3, ust. 1 pkt 17.*

[7] M. Łuniewska, Ekonometria finansowa Analiza Rynku Kapitałowego, Warszawa: Wydawanictwo Naukowe PWN SA, 2008.

[8] A. Laidi, Międzyrynkowa analiza kursów walutowych, Warszawa: Wydawnictwo Linia Sp. z o.o., 2012.

[9] M. T. Leung, H. Daouk i A.-S. Chen, „Forecasting stock indices: a comparison of classification and level estimation models," *International Journal of Forecasting,* tom 16, pp. 173-190, 2000.

[10] K. Senthamarai Kannan, P. Sailapathi Sekar, M. Mohamed Sathik i P. Arumugam, „Financial Stock Market Forecast using Data Mining Techniques," *Proceedings of the International MultiConference of Engineers and Computer Scientists,* tom 1, 2010.

[11] J. J. Murphy, Analiza techniczna rynków finansowych, Warszawa: WIG-Press, 1999.

[12] J. Brzeszczyński i R. Kelm, Ekonometryczne modele rynków finansowych. Modele kursów giełdowych i kursów walutowych., Warszawa: WIG-Press, 2002.

[13] J. C. Ritchie, Analiza fundamentalna, Warszawa: WIG-Press, 1997.

[14] I. H. Witten i E. Frank, Data Mining Practical Machine Learning Tools and Techniques, Second Edition, San Francisco: Elsevier Inc., 2005.

[15] E. Hajizadeh, H. D. Ardakani i J. Shahrabi, „Application of data mining techniques in stock markets: A survey," *Journal of Economics and International Finance,* tom 2, nr 7, pp. 109-118, 2010.

[16] W. Ian H. i F. Eibe, Data Mining Practical Machine Learning Tools and Techniques, San Fransisco: Elsevier Inc., 2005.

[17] M. Subha i S. Thirupparkadal Nambi, „Classification of Stock Index movement using k-Nearest Neighbours (k-NN) algorithm,” *WSEAS TRANSACTIONS ON INFORMATION SCIENCE and APPLICATIONS,* tom 9, nr 9, pp. 261-270, 2012.

[18] T. Fu, C. Shuo i W. Chuanqi, „Hong Kong Stock Index Forecasting”.

[19] C. Hargreaves i Y. Hao, „Prediction of Stock Performance Using Analytical Techniques,” *Journal of emerging technologies in web intelligence,* tom 5, nr 2, pp. 136-142, 2013.

[20] M. D. Rechenthin, „Machine-learning classification techniques for the analysis and prediction of high-frequency stock direction,” University of Iowa, Iowa, 2014.