POLITECHNIKA BIAŁOSTOCKA

WYDZIAŁ Informatyki

Katedra Oprogramowania

PRACA DYPLOMOWA MAGISTERSKA

TEMAT: **Krótkoterminowa prognoza kierunku zmiany kursów walut z wykorzystaniem uczenia maszynowego.**

WYKONAWCA: Maciej Ziniewicz
Imię i nazwisko

PODPIS: .....

PROMOTOR: dr inż. Jerzy Krawczuk
Imię i nazwisko

PODPIS: .....

BIAŁYSTOK 2017 ROK

Karta dyplomowa

POLITECHNIKA BIAŁOSTOCKA Wydział Informatyki Katedra Oprogramowania	Studia stacjonarne II stopnia	Nr albumu studenta 90563 Rok akademicki 2016/2017 Kierunek studiów Informatyka Specjalność Inteligentne technologie internetowe

Maciej Ziniewicz

TEMAT PRACY DYPLOMOWEJ:

Krótkoterminowa prognoza kierunku zmiany kursów walut z wykorzystaniem uczenia maszynowego.

Zakres pracy:

1. Poszukiwanie optymalnego algorytmu prognozy zmian wybranych par walutowych.
2. Symulacja gry na opcjach binarnych z wykorzystaniem zbudowanego modelu.

Słowa kluczowe (max 5): uczenie maszynowe, java, giełda walutowa, prognoza, analiza techniczna

..... *Imię i nazwisko, stopień/ tytuł promotora - podpis*	..... *Imię i nazwisko kierownika katedry - podpis*

..... .*Data wydania tematu pracy dyplomowej - podpis promotora*	30.09.2017 *Regulaminowy termin złożenia pracy dyplomowej*	.... *Data złożenia pracy dyplomowej - potwierdzenie dziekanatu*

..... *Ocena promotora*	..... *Podpis promotora*

..... *Imię i nazwisko, stopień/ tytuł recenzenta*	..... *Ocena recenzenta*	.... *Podpis recenzenta*

Thesis topic:
Short-term forecast of the direction of change in exchange rates using machine learning.

SUMMARY

The aim of the study is to investigate and select the optimal machine learning algorithm and perform a simulation of the game on the binary options. Binary Options is a new stock market instrument based on speculations whether the price of a given stock market asset will increase or fall within a certain time interval. In this case market asset is an currency pair. According to the assumption of binary options, in research, machine learning algorithms will have the task to speculate the direction of price movement within a certain time interval.

I chose this thesis topic because of the interest in stock market investments, where I have basic experience and by writing this thesis I wanted to explore the principles of technical analysis and the construction of the indicators used. In addition, I find machine learning to be an extremely interesting. Using it to build artificial intelligence and analysis of large data sets, can become one of the core components of future computing. Therefore, another motivation to write a work as strong as the desire to learn technical analysis was to work with machine learning.

The work is divided into 6 chapters. The second chapter will provide a theoretical introduction to machine learning and basic stock market information. Chapter three shows the tools that were used to generate data sets and to perform research. The next, fourth chapter describes the method of research by explaining it step by step. In the penultimate fifth chapter, research and simulation results are presented. The last sixth chapter is a summary, describes the conclusions of the thesis and suggests further research directions.

Maciej Ziniewicz .
imię i nazwisko studenta
90563
nr albumu
Informatyka stacjonarne
kierunek i forma studiów
dr inż. Jerzy Krawczuk
promotor pracy dyplomowej

Białystok, dnia……

OŚWIADCZENIE

Przedkładając w roku akademickim 2016/2017. Promotorowi dr inż. Jerzemu Krawczukowi pracę dyplomową pt.: Krótkoterminowa prognoza kierunku zmiany kursów walut z wykorzystaniem uczenia maszynowego.,

oświadczam, że:

1) praca dyplomowa stanowi wynik samodzielnej pracy twórczej,
2) wykorzystując w pracy dyplomowej materiały źródłowe, w tym w szczególności: monografie, artykuły naukowe, zestawienia zawierające wyniki badań (opublikowane, jak i nieopublikowane), materiały ze stron internetowych, w przypisach wskazywałem ich autora, tytuł, miejsce i rok publikacji oraz stronę, z której pochodzą powoływane fragmenty, ponadto w pracy dyplomowej zamieściłem bibliografię,
3) praca dyplomowa nie zawiera żadnych danych, informacji i materiałów, których publikacja nie jest prawnie dozwolona,
4) praca dyplomowa dotychczas nie stanowiła podstawy nadania tytułu zawodowego, stopnia naukowego, tytułu naukowego oraz uzyskania innych kwalifikacji,
5) treść pracy dyplomowej przekazanej do dziekanatu Wydziału Informatyki jest jednakowa w wersji drukowanej oraz w formie elektronicznej,
6) jestem świadomy/a, że naruszenie praw autorskich podlega odpowiedzialności na podstawie przepisów 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.), jednocześnie na podstawie przepisów ustawy z dnia 27 lipca 2005 r. Prawo o szkolnictwie wyższym (t. j.: Dz. U. z 2012 r. poz. 572 z późn. zm.) stanowi przesłankę wszczęcia postępowania dyscyplinarnego oraz stwierdzenia nieważności postępowania w sprawie nadania tytułu zawodowego,
7) udzielam Politechnice Białostockiej nieodpłatnie licencji na korzystanie z pracy dyplomowej w celu realizacji przeprowadzenia procedury antyplagiatowej przyjętej w Uczelni oraz na przekazanie pracy do Ogólnopolskiego Repozytorium Prac Dyplomowych, jak również udostępnianie i przechowywanie jej w Bibliotece Politechniki Białostockiej przez okres 50 lat od obrony pracy dyplomowej.

……
czytelny podpis studenta

Na podstawie art. 24 ust. 1 ustawy z dnia 29 sierpnia 1997 roku o ochronie danych osobowych (j.t. z 2014 r., poz. 1182 z późn. zm.) informuję, że administratorem danych jest Politechnika Białostocka, ul. Wiejska 45A, 15-351 Białystok. Dane będą przetwarzane w celach realizacji procedury antyplagiatowej przyjętej w Politechnice Białostockiej i nie będą udostępniane odbiorcom danych w rozumieniu art. 7 pkt 6 ustawy o ochronie danych osobowych. Osobie, której dane dotyczą, przysługuje prawo dostępu do treści swoich danych oraz ich poprawiania. Podanie danych jest obowiązkowe (art. 167b ustawy z dnia 27 lipca 2005 roku Prawo o szkolnictwie wyższym j.t. Dz.U. z 2012 r., poz. 572 z późn. zm.).

Spis treści

1. Wstęp

W ciągu ostatnich lat znacznie wzrosło zainteresowanie algorytmami uczenia maszynowego oraz wykorzystaniem ich do analizowania dużych zbiorów danych. Obecnie zagadnienia uczenia maszynowego zdobywają uznanie w ogromnej ilości dziedzin i są używane do prognozy pogody, katalogowania i dodawania nowych produktów w sklepach, rozpoznawania mowy, kierowania pojazdami czy rozwoju robotyki. Dziedzina ta świetnie nadaje się do analizy, badania i poszukiwania zależności w dużych ilościach danych.

Inwestowanie na giełdzie może przynieść bardzo duże zyski, które kuszą ludzi do podejmowania różnych metod i strategii tak by prognozować przyszłe zachowania ceny. Jedną z takich metod jest użycie uczenia maszynowego, które wydaje się być bardzo dobrym narzędziem do celów tego typu. Rynek dostarcza bardzo dużą ilość pozornie niemających związku ze sobą informacji, które są trudne do połączenia nawet dla człowieka. Za pomocą algorytmów uczenia maszynowego można te informacje analizować, wspomagać procesy decyzyjne, a w najlepszym przypadku podejmować inwestycje. Niestety rynek charakteryzuje się dużą ilością szumów i niestabilnością przez co prognoza jest zadaniem bardzo trudnym.

Celem pracy jest badanie i wybór optymalnego algorytmu uczenia maszynowego oraz przeprowadzenie symulacji gry na opcjach binarnych. Opcje binarne to nowy instrument giełdowy polegający na spekulowaniu czy cena danego aktywa giełdowego w tym przypadku pary walutowej wzrośnie czy spadnie w określonym przedziale czasowym. Zgodnie z założeniem opcji binarnych algorytmy uczenia maszynowego, w badaniach będą miały za zadanie spekulację kierunku ruchu ceny w określonym przedziale czasowym, natomiast dokładna wartość ceny nie będzie prognozowana.

Temat ten wybrałem ze względu na zainteresowanie inwestycjami giełdowymi, w których mam podstawowe doświadczenie, a poprzez pisanie tej pracy chciałem zgłębić zasady działania analizy technicznej oraz budowę użytych wskaźników. Dodatkowo uważam uczenie maszynowe za niezwykle interesującą dziedzinę, która wykorzystana do budowania sztucznej inteligencji i analizy dużych zbiorów danych może stać się jednym z podstawowych elementów przyszłej informatyki. Dlatego też kolejną motywacją do podjęcia tego tematu, równie silną jak chęć poznania analizy technicznej, była praca z zagadnieniami uczenia maszynowego i wykorzystania ich do prognozy kierunków kursów.

Praca podzielona jest na 6 rozdziałów. W rozdziale drugim opisano wstęp teoretyczny wprowadzający w zagadnienia uczenia maszynowego oraz podstawowych informacji o giełdzie. Rozdział trzeci przedstawia narzędzia jakie zostały użyte do wygenerowania zbiorów danych oraz do przeprowadzenia badań. Kolejny, czwarty rozdział opisuje metodę wykonywanych badań wyjaśniając w kolejności każdy krok. W przedostatnim piątym rozdziale zaprezentowane zostały wyniki badań oraz symulacji. Ostatni szósty rozdział stanowi podsumowanie, opisuje wnioski płynące z pracy magisterskiej i sugeruje dalsze kierunki badań.

2. Prognoza rynków

Poniższy podrozdział stanowi wprowadzenie teoretyczne do pojęć i technik jakie wiążą się z predykcją rynków giełdowych. Pierwszy podrozdział 2.1 jest wprowadzeniem teoretycznym do rynku Forex, opisuje czym jest ten rynek oraz czym są pary walutowe. Kolejny podrozdział 2.2 przedstawia pojęcia analizy fundamentalnej, natomiast podrozdział 2.3 opisuje analizę techniczną oraz techniki z nią związane. Ostatni podrozdział przedstawia dziedzinę jaką jest uczenie maszynowe, podstawowe pojęcia jakie są niezbędne przy pracy z uczeniem maszynowym oraz wybrane algorytmy.

2.1 Rynek Forex

Foregin Exchange Market częściej nazywany Forex (ang. Foregin Exchange) jest to największy na świecie rynek walutowy, gdzie rządy, banki centralne, maklerzy i inni inwestorzy wykonują operacje wymiany walut. Początkowo tylko wielkie i bogate podmioty handlowe miały możliwość użytkowania rynku Forex jednak z przyjściem i popularyzacją Internetu możliwość tą otrzymali również przeciętni ludzie, którzy mogą teraz podejmować decyzje inwestycyjne nie wstając od komputera. Rynek ten będzie funkcjonował tak długo jak ludzie będą używali walut. Otwarty 24 godziny na dobę, a większość platform inwestycyjnych umożliwiających wykonywanie inwestycji dla zwykłych osób jest czynna od 22:00 GMT w niedzielę do 22:00 GMT piątek. Handel odbywa się tam w różnych sesjach handlowych: azjatyckiej, amerykańskiej i europejskiej przez cały czas pomijając weekendy oraz ważne święta. Każda waluta na Forex jest oznaczona specjalnym trzyliterowym kodem według norm ISO przykładowo euro to EUR a dolar amerykański to USD. Notowania są mierzone w dokładności do 1 pip (ang. Price Interest Point) i jest to najmniejsza wartość zmiany jaka może nastąpić w cenie danej waluty. Dla większości walut pip wynosi 0,0001 część całości np. 3,5212 [15].

2.1.1 Pary walutowe

Aby wykonywać transakcje handlowe z użyciem walut na rynku Forex należy znać ich wartość, którą określa się w porównaniu do innej waluty. Przykładowo wartość euro do złotówki wynosząca 4.23 oznacza, że 1 euro jest wart 4.23 złotego, czy wartość euro do dolara 1.2 oznacza, że jeden euro jest wart 1.2 dolara amerykańskiego. Porównania

w taki sposób różnych walut tworzą pary walutowe EUR/PLN o kursie 4.23 czy jedna z najpopularniejszych EUR/USD o kursie 1.2. Walutę pierwszą w parze walutowej nazywamy walutą bazową, natomiast drugą walutą kwotowaną.

Wyróżniane są trzy podstawowe typy walut [16]:

a) **Główne** – są wszystkie pary odnoszące się do dolara amerykańskiego po jednej ze stron, kwotowanej lub bazowej. Główne pary z racji, że są najczęściej używane wyróżniają się wysoką płynnością oraz niskimi kosztami transakcji. Transakcje dotyczące par z tej kategorii stanowią większość transakcji na rynku Forex.

Para	Kraje	Udział % w Forex
USD/EUR	Stany Zjednoczone/Strefa Euro	24.1
USD/JPY	Stany Zjednoczone/Japonia	18.3
USD/GBP	Stany Zjednoczone/Wielka Brytania	8.8
USD/AUD	Stany Zjednoczone/Australia	6.8
USD/CAD	Stany Zjednoczone/Kanada	3.7
USD/CHF	Stany Zjednoczone/ Szwajcaria	3.4

Tabela 2.1 Główne pary walutowe dane z 2013
Źródło: opracowanie własne na podstawie danych z Bank for international settlement [10]

b) **Krzyżowe** – są to pary które nie odnoszą się do dolara amerykańskiego. Pary krzyżowe nie są tak popularne jak główne przez co koszty handlu nimi są większe.

Para	Kraje
EUR/AUD	Strefa Euro / Kanada
EUR/JPY	Strefa Euro / Japonia
EUR/NZD	Strefa Euro / Nowa Zelandia
CHF/JPY	Szwajcaria / Japonia
GBP/AUD	Wielka Brytania / Kanada

Tabela 2.2 Krzyżowe pary walutowe
Źródło: Opracowanie własne

c) **Egzotyczne** - są to pary walutowe składające się z jednej z głównych bardziej znaczących walut do walut które są mniej znaczące ekonomicznie w skali globalnej. Handel z użyciem tych par walutowych stanowi najmniejszą część rynku Forex.

Para	Kraje
EUR/PLN	Strefa Euro / Polska
EUR/HKD	Strefa Euro / Hongkong
USD/SGD	Stany Zjednoczone / Singapur
USD/ZAR	Stany Zjednoczone / Afryka Południowa

Tabela 2.3 Egzotyczne pary walutowe
Źródło: Opracowanie własne

2.2 Opcje binarne

Opcje binarne to nowy rodzaj instrumentu finansowego na Polskim rynku. Są one znane inwestorom giełdowym od dłuższego czasu, jednak od niedawna są powszechnie dostępne dla każdego. Charakteryzują się one prostymi zasadami użycia i inwestycji za pośrednictwem dostępnych platform. Inwestor nie kupuje walut, tylko nabywa opcję dotyczącą waluty. Nabyta opcja ma określony termin wygaśnięcia zgody z ustalonym przedziałem czasowym. Zadaniem inwestora jest przewidzenie czy wartość waluty, której dotyczy opcja będzie wyższa czy niższa w momencie wygaśnięcia opcji, czyli z końcem ustalonego przedziału czasowego, od chwili jej zakupienia. Jeżeli decyzja okazuje się trafna to inwestor zarabia. Zysk z wygrywającej opcji mieści się zazwyczaj w zakresie od 60% do 85% jednak zdarzają się brokerzy, którzy oferują nawet 91% stopę zwrotu. W przypadku, gdy opcja nie jest wygrywająca i kierunek ruchu ceny nie był zgodny z decyzją podjętą przez inwestora, pieniądze zainwestowane w opcje są tracone. Różnicą między opcjami binarnymi a forexem jest to że, nawet ruch o jeden pips może przynieść zysk, jak i starte z nabytej opcji. Oraz inwestor z góry jest świadomy tego, ile ryzykuje. Ponieważ w opcjach jedyna kwota jaką się ryzykuje to kwota za jaką została kupiona opcja, natomiast w przypadku forex ryzykuje się całym włożonym kapitałem co w przypadku złego zarządzania portfelem oraz niezabezpieczeniu otwartych opcji może skutkować stratą całości kapitału.

Rysunek 2.1 Widok aplikacji IQ Options brokera opcji binarnych
Źródło: https://xbinop.com/pl/opinia/iqoption/ (stan na 16.09.2017)

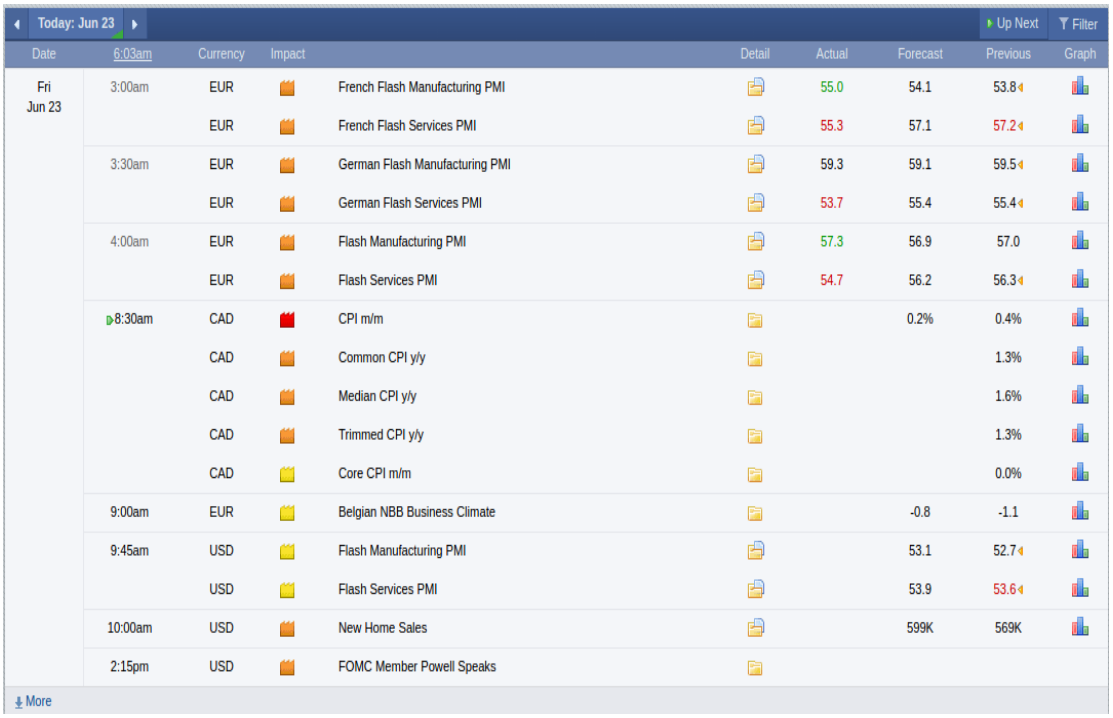
Powyższy obrazek przedstawia widok z aplikacji dostarczonej przez najpopularniejszego brokera opcji binarnych IQ Options. Umożliwia ona zdefiniowanie kwoty jaką chce się zainwestować w opcję, wybór przedziału czasowego, określenie za pomocą przycisków „CALL" i „PUT" kierunku ruchu ceny wybranej waluty, oraz prezentuje stopę zwrotu w tym przypadku wynoszącą 90%. Wybór „CALL" jest to ruch ceny w górę natomiast wybór „PUT" jest to ruch ceny w dół względem ceny w momencie nabycia opcji. Dodatkowymi narzędziami jakie są oferowane to wskaźniki analizy technicznej, informacje o wydarzeniach oraz trzy typy wykresów.

Opcje binarne są interesującym instrumentem finansowym szczególnie dla początkujących inwestorów. Charakteryzują się prostymi zasadami oraz wysoką stopą zwrotu w przypadku wygranej.

2.3 Analiza fundamentalna

Jest to jedna z technik analizy rynków. Głównym zadaniem analizy fundamentalnej jest śledzenie i analiza czynników zewnętrznych mających wpływ na popyt i podaż takich jak: decyzje banków, przemowy ważnych polityków, prezesów dużych firm, publikacje raportów. Celem jest by na podstawie analizy takich informacji zareagować i wykonać odpowiednią transakcje zanim informacje wpłyną na rynek. Wszelkiego rodzaju nowe wiadomości są ważne z punktu widzenia analizy fundamentalnej,

ponieważ mają potencjalny wpływ na rynek [5]. Tego rodzaju analizy są stosowane głownie w długoterminowych inwestycjach oraz częściej to pojęcie jest spotykane przy rynkach akcyjnych, gdy określa się wartość spółek. Nie znaczy to jednak że analiza fundamentalna nie jest stosowana na rynku Forex. Podstawowym narzędziem jakie jest używane podczas inwestycji na giełdzie walutowej jest kalendarz ekonomiczny, gdzie można odnaleźć informacje na temat wydarzeń mających wpływ na poszczególne waluty.



Rysunek 2.2 Kalendarz ekonomiczny
Źródło: www.forexfactory.com (stan na 8.07.2017)

Na powyższym obrazku widać przykładowy kalendarz ekonomiczny dla dnia 23 czerwca 2017r. Wydarzenia podzielone są w nim początkowo ze względu na godzinę o której występują. Dla każdej godziny przyporządkowana jest lista wydarzeń, które wtedy będą miały miejsce. Każde wydarzenie ma podaną walutę na jaką będzie miało wpływ i określoną jego siłę, w tym przypadku trzema różnymi kolorami (w innych kalendarzach może być to inaczej prezentowane, ale zazwyczaj jest to prezentacja w trzech stopniach wpływu):

- Żółty – najmniejszy wpływ
- Pomarańczowy – średni wpływ
- Czerwony – duży wpływ

oraz podane są takie informacje jak rodzaj wydarzenia, dodatkowe szczegóły, realny wpływ, jeżeli wydarzenie już miało miejsce, prognozę, oraz poprzedni wpływ.

2.4 Analiza techniczna

Analiza techniczna jest to kolejna z podstawowych technik analizy rynków, a jej definicja opisuje, że jest to „badanie zachowań rynku, przede wszystkim przy użyciu wykresów, którego celem, jest przewidywanie przyszłych trendów cenowych" [4]. Przede wszystkim analiza techniczna polega na odczytywaniu informacji z wykresów instrumentów giełdowych, którymi w tym przypadku będą to waluty. Na podstawie obserwacji zachowań ceny m.in: nagłych skoków, utworzonych formacji prognozuje się przyszły trend. Cała koncepcja analizy technicznej opiera się o trzy przesłanki [4, 14]:

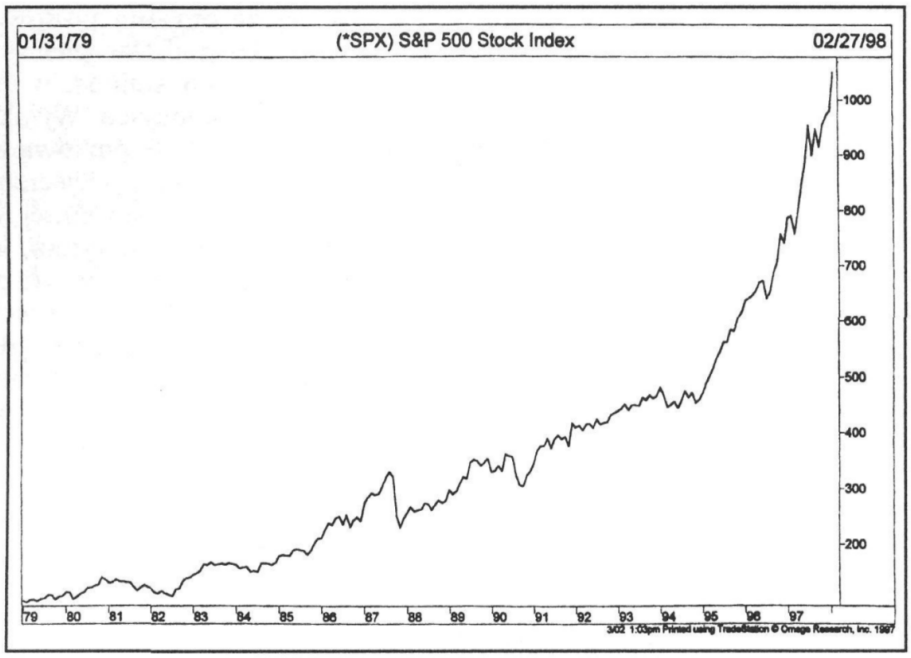
- Rynek dyskontuje wszystko
- Ceny podlegają trendom
- Historia się powtarza

a) Rynek dyskontuje wszystko

Ta przesłanka jest podstawą analizy technicznej z której wynikają kolejne. Zakłada ona, że wszystkie czynniki od fundamentalnych po psychologię rynku mające wpływ na cenę znajdują odbicie w jej zachowaniu. Wynika z tego, że do określenia przyszłego zachowania rynku wystarczy badanie zachowania cen, i nie trzeba badać każdego z czynników oddzielnie, ponieważ wykres odzwierciedla je wszystkie na raz. Zgodnie z tą tezą, jeżeli cena rośnie to popyt przewyższa podaż a sytuacja na rynku musi być sprzyjająca wzrostom, natomiast jeżeli cena spada to analityk wychodzi z założenia, że podaż przewyższa popyt a sytuacja rynkowa sprzyja spadkom.

b) Ceny podlegają trendom

Kolejne podstawowe założenie analizy technicznej mówi, że ceny podlegają trendom krótko, średnio, długo terminowym. Rynek wykazuje większą tendencję do kontynuowania trendu niż do jego zmiany. Innymi słowy duża część analizy technicznej sprowadza się do tego by badać trendy oraz wykrywać ich zmiany w wystarczająco wczesnej fazie co pozwoli dokonywać zyskowne transakcje.



Rysunek 2.3 Trend wzrostowy
Źródło: Murphy, Analiza techniczna rynków finansowych, WIG-PRESS Warszawa 1999, s 4

c) Historia się powtarza

Analiza techniczna w dużym stopniu powiązana jest z badaniem ludzkiej psychiki, ponieważ ludzie mają tendencję do powtarzalności swoich zachowań. Formacje świecowe znane są od lat, odzwierciedlają one zachowania ludzkie. Powtarzalność zachowań ceny jest podstawą tej tezy, ponieważ jeżeli formacje cenowe sprawdziły się w przeszłości to zakłada się, że będą skuteczne również w przyszłości, dlatego analiza techniczna używa narzędzi do badań przeszłości, aby określić zachowanie ceny w przyszłości.

2.4.1 Świece japońskie

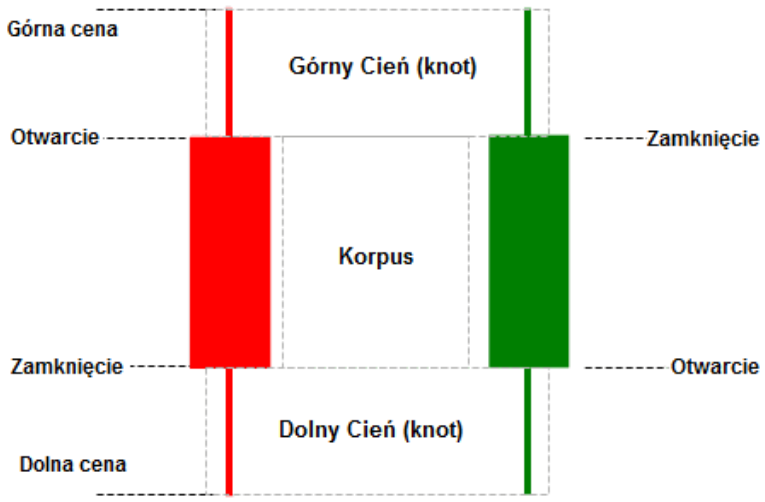
Świece japońskie jest to jedna z trzech najpopularniejszych metod przedstawiania wykresów rynków. Jak sama nazwa wskazuje świece te pochodzą z Japonii a pierwsze informacje o ich użyciu sięgają XVIII w, gdzie były używane przez japońskiego kupca ryżu Munehisa Homma. Który w późniejszym czasie został konsultantem finansowym rządu japońskiego oraz honorowo mianowany samurajem, a jego wskazówki handlowe i książka wpłynęły na późniejsze ukształtowanie się obecnie używanych świec [8].

Wykres świecowy oraz słupkowy dają większą ilość informacji, w porównaniu do innego również często używanego wykresu liniowego.

Rysunek 2.4 Rodzaje wykresów
Źródło: opracowanie własne

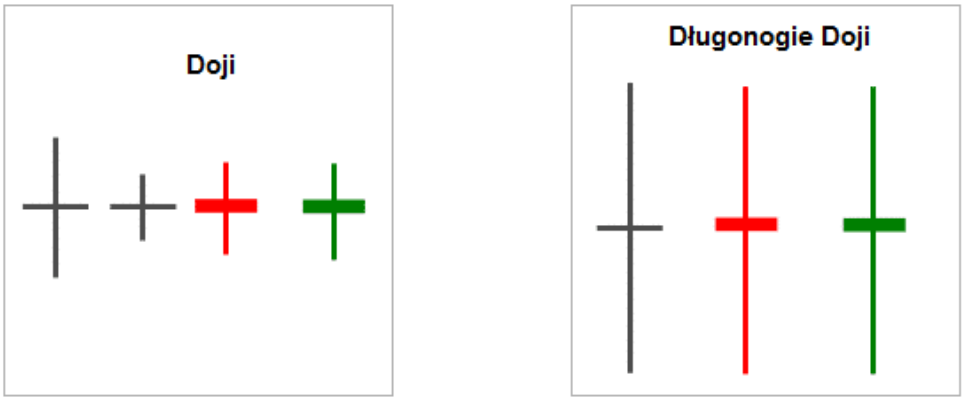
Tak jak widać na powyższym rysunku (rys. 2.3) wykres liniowy dostarcza najmniejszą ilość informacji dla inwestora. Łączy on ceny zamknięcia z każdego przedziału czasowego to znaczy, że inwestor otrzymuje informacje o cenie otwarcia, czyli cenie początkowej w określonym przedziale czasowym oraz cenie zamknięcia tj. cenie końcowej. Wykres słupkowy dostarcza nam dodatkowo informacje o maksymach jakie wystąpiły podczas ruchu ceny w określonym przedziale czasowym, czyli wartość maksymalną i minimalną jakie wystąpiły w tym czasie. Podobnie jak wykresy słupkowe wykresy świecowe dostarczają informacje o cenie otwarcia, zamknięcia oraz maksymalnej i minimalnej wartości ceny w przedziale czasowym.

Świece japońskie składają się z korpusu obejmującego zakres od ceny otwarcia do ceny zamknięcia. Kolor świecy również ma znaczenie, standardowo są to kolory biały zastąpiony tutaj zielonym oraz czarny zastąpiony czerwonym. Jeżeli kolor świecy jest czerwony to cena zamknięcia jest niższa niż cena otwarcia i jest to tak zwana świeca spadkowa. Natomiast jeżeli świeca jest zielona to cena zamknięcia jest wyższa niż cena otwarcia i jest to świeca wzrostowa. Dodatkowymi informacjami dostarczanymi są przez świece są jej cienie, a informują one o ekstremach ceny występujących podczas trwania okresu reprezentowanego przez świecę. Cień górny informuje o cenie maksymalnej, a cień dolny informuje o cenie minimalnej.



Rysunek 2.5 Świeca spadkowa oraz wzrostowa
Źródło: https://comparic.pl/swiece-japonskie-podstawy/ (dostęp na 08.07.2017)

Wygląd pojedynczych świec może stanowić dodatkową informację. Najbardziej podstawowym oraz niemniej ważnym i znaczącym wzorcem jest tzw. „doji" czyli świeca o tej samej cenie otwarcia i zamknięcia lub o bardzo małym korpusie, sygnalizuje ona niezdecydowanie rynku i możliwą zmianę trendu. Doji mogą różnić się długością cienia.



Rysunek 2.6 Wzorzec doji
Źródło: https://comparic.pl/swiece-japonskie-podstawy/ (dostęp na 08.07.2017)

Dla inwestora również istotną wiadomością jest sposób ułożenia świec kolejno po sobie. Charakterystyczny i powtarzający się układ wielu świec następujących po sobie nazywany jest formacją [8].

2.4.2 Analiza trendu

Pojęcie trendu jest absolutną podstawą przy analizie technicznej, ponieważ wszystkie narzędzia, którymi posługuje się inwestor służą do badania trendów. Przy inwestowaniu

można często spotkać się z powiedzeniami, że walka z trendem, czyli inwestycje w kierunku przeciwnym niż idzie trend nie maja sensu, dlatego zrozumienie pojęcia trendu jest tak ważne. Trend można zdefiniować jako „kierunek, jaki przyjmują szczyty i dołki” [4], co znaczy, że trend wzrostowy można zdefiniować jako szczyty i dołki kolejno występujące po sobie coraz wyżej o coraz wyższych wartościach, natomiast trend spadkowy byłyby to kolejno występujące po sobie szczyty i dołki położone coraz niżej.

Rysunek 2.7 Trend wzrostowy i spadkowy
Źródło: opracowanie własne

Istnieje jeszcze trend boczny, czyli szczyty i dołki, które następują po sobie horyzontalnie. W taki sposób, że różnice pomiędzy ich wartościami są nieznaczne, i cena nie posiada szczególnego kierunku zmiany. Taka sytuacja również określana jest czasem jako brak trendu, ponieważ cena nie podąża ani do góry, ani do dołu.

Rysunek 2.8 Trend boczny
Źródło: opracowanie własne

Trendy rozróżnia się nie tylko ze względu na kierunek, ale również na rodzaj. Mianowicie trend określając ze względu na rodzaj może być główny, średnio lub krótko okresowy. Trend główny jest to najdłuższy trend i żeby określić trend jako główny według teorii Dowa trwać on musi co najmniej rok. Trend średnio okresowy jest częścią trendu głównego, tak jak trend krótko okresowy jest częścią trendu średniookresowego i głównego. Generalnie w większych trendach występują mniejsze trendy korygujące jednak wyróżniane są najczęściej trzy powyższe. Przykładowo, gdy trend główny jest wzrostowy, może w nim wystąpić krótszy trend korygujący średnio okresowy spadkowy, który obniży cenę. Następnie jak trend korygujący skończy się cena będzie nadal poruszała się zgodnie z trendem głównym [4,7].

2.4.3 Wskaźniki

Wskaźniki używane są jako podstawowe narzędzie analizy technicznej obok interpretacji wykresów, świec czy występujących formacji. Jest to statystyczne podejście do analizy technicznej oraz znacznie wspomaga proces decyzyjny podczas klasycznej subiektywnej analizy technicznej, czyli interpretacji tego co inwestor widzi na wykresach. Wskaźniki generują dodatkowe informacje odnośnie sytuacji na rynku oraz pomagają w identyfikacji stanów w jakich znajduje się rynek. Są to narzędzia analizujące dane trudno widoczne na pierwszy rzut oka tj. przepływy pieniężnie, trendy, zmienność i dynamikę, oraz reprezentują je za pomocą wykresów, na podstawie których możliwe jest generowanie sygnałów. Wskaźniki mimo to że mają wspierać inwestora w procesie decyzyjnym mogą przynieść efekt odwrotny, dlatego nie należy traktować ich jako podstawę do podejmowania decyzji. Nie zaleca się również korzystać ze zbyt dużej ilości wskaźników, ponieważ może to wprowadzać w błąd.

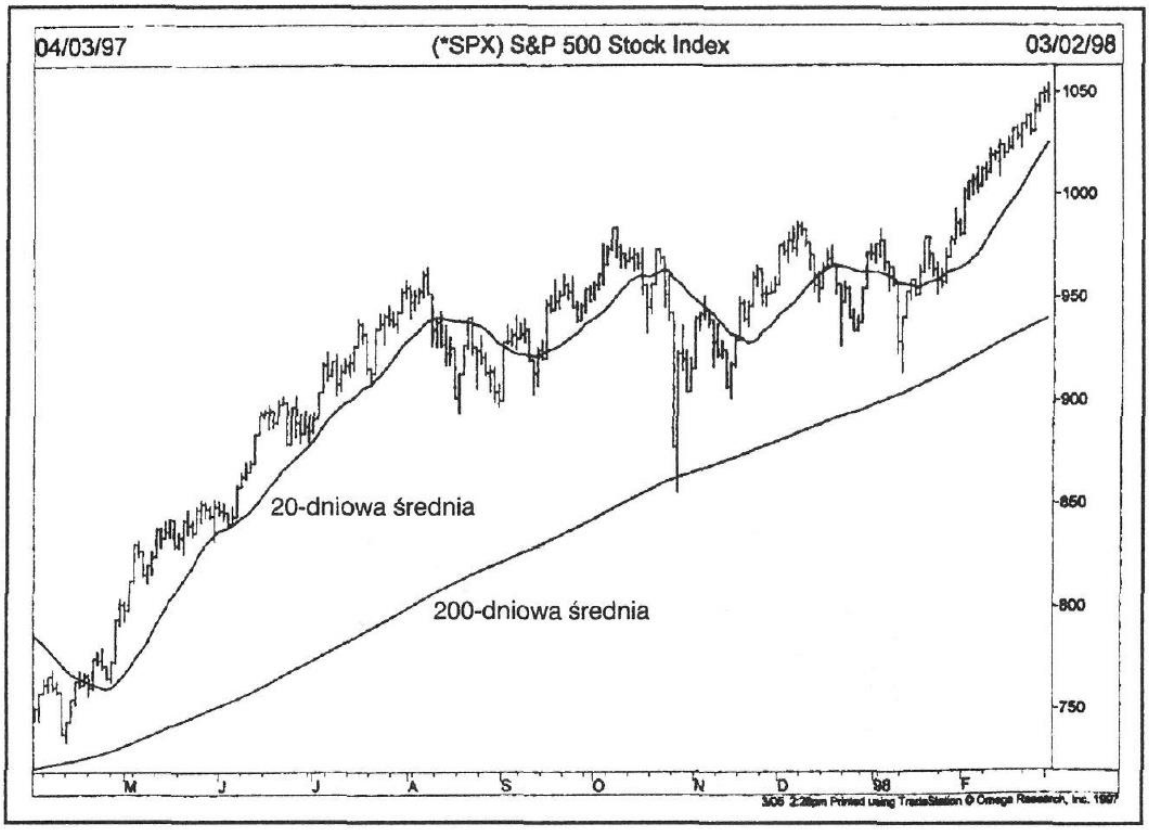
Wskaźniki możemy podzielić na dwa typy [14]:

1. Wiodące (ang. leading) – wskaźniki poprzedzające ruchy ceny oraz próbujące je przewidzieć. Przydają się podczas trendów bocznych, gdy rynek nie ma widocznych trendów.
2. Opóźnienia (ang. lagging) – wskaźniki podążające za ruchem ceny, śledzące ją. Służą najczęściej jako potwierdzenia wspomagają decyzje. Przydatne podczas trendów.

Poza ogólną klasyfikacją wskaźniki można grupować na podstawie tego co wskazują, np.: zmienność ceny, siła ceny, średnie ruchome, popyt, wolumen, itd. Na podstawie wskaźników można wywnioskować lub wygenerować sygnał zakupu/sprzedaży.

2.4.3.1 Średnie ruchome

Średnia ruchoma (ang. MA – Moving Average) jest jednym z najbardziej uniwersalnych i popularnych wskaźników stosowanych w analizie technicznej. Cechuje się prostą budową dzięki czemu łatwo go zaprogramować oraz sygnały trendu dawane przez średnie są precyzyjne. Jak sama nazwa wskazuje wskaźnik ten jest średnią notowań z określonego przedziału czasowego. Przykładowo, aby uzyskać średnią dwudziestodniową, należy zsumować ze sobą ceny z ostatnich dwudziestu dni oraz podzielić wynik przez dwadzieścia. Następnym członek nazwy jest „ruchoma", oznacza to uwzględnienie ceny tylko z podanego okresu, w tym przypadku z dwudziestu dni. Każdego kolejnego dnia do sumy dwudziestu dni dodawane jest nowe notowanie, a odejmowane ostanie. W taki sposób średnia przesuwa się za ceną.

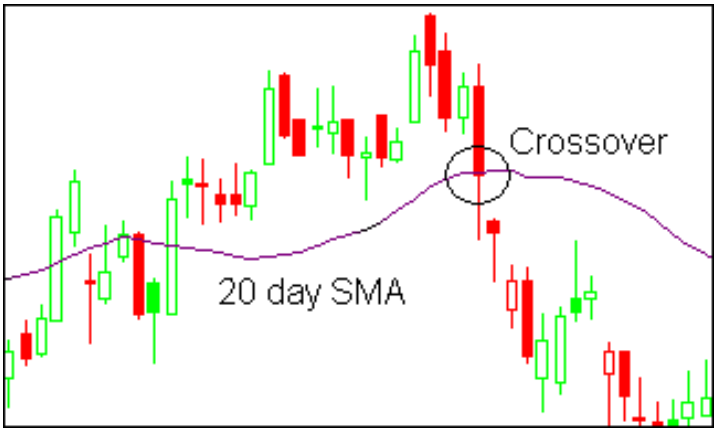


Rysunek 2.9 Porównanie średniej ruchomej 20- i 200-dniowej.
Źródło: Murphy, Analiza techniczna rynków finansowych, WIG-PRESS Warszawa 1999, s 175

Średnie są narzędziami badania już istniejącego trendu, za którym podążają. Ich zadaniem jest sygnalizacja wystąpienia nowego trendu lub wykrycie zmiany obecnego. Średnie krótkoterminowe są bardziej czułe niż średnie długo terminowe, oznacza to konieczność dostosowania dostarczanego do wskaźnika przedziału notowań do analizowanego okresu przedziału czasowego. Można użyć do analizy krótkoterminowej średniej kroczącej dwustudniowej jednak nie przyniesie to dobrych wyników.

Przedstawiony wyżej sposób obliczeń dotyczył prostej średniej ruchomej (ang. SMA – Simple Moving Average) i nie jest jedyną metodą obliczania średnich kroczących. Innymi popularnymi sposobami obliczeń są średnie ważone (ang. WMA – Weighted Moving Average) oraz średnie wykładnicze (ang. EMA – Exponential Moving Average). Średnie ważone polegają na nadawaniu wag kolejnym notowaniom użytym do obliczeń średniej tak by większa waga była przywiązywana wprost proporcjonalnie do wzrostu cen. A następnie zsumowane wagi są dzielone przez sumę mnożników. Średnia wykładnicza jest modyfikacją średniej ważonej i nadaj ona większe wagi bardziej aktualnym cenom [4].

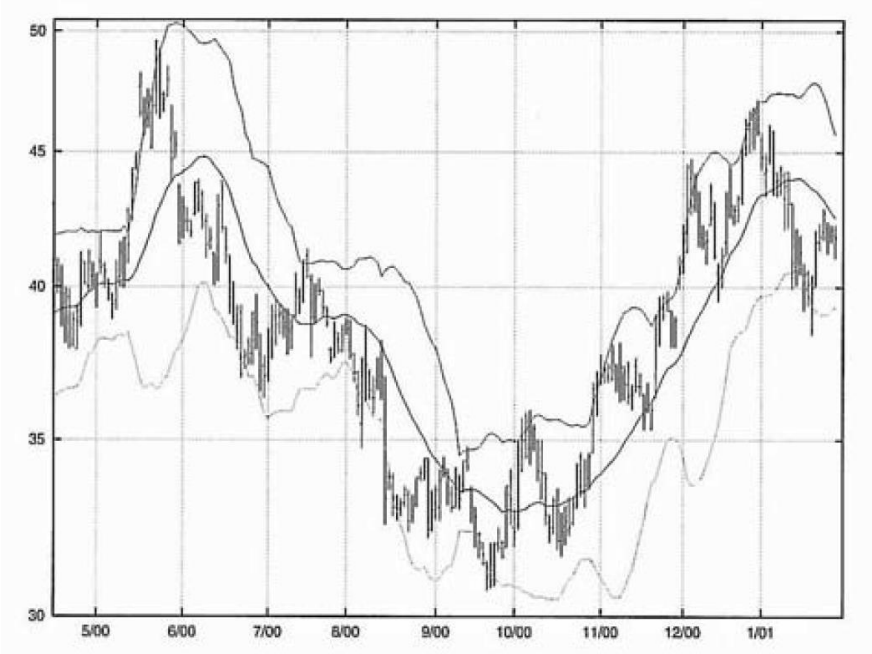
Sygnałami otrzymywanymi ze średnich są przecięcia linii. Cena częściej wacha się i idzie w jakimś kierunku rysując wykres nieregularnie niż po linii prostej. Przy takim ruchu, gdy cena przetnie linie rysowaną przez wskaźnik opóźnienia może być to interpretowane jako sygnał odwrócenia trendu [14]. Poniżej przykład przecięcia linii dwudziestodniowej średniej kroczącej.



Rysunek 2.10 Przecięcie linii średniej kroczącej
Źródło: http://www.investopedia.com/terms/c/crossover.asp (dostęp 11.07.2017)

2.4.3.2 Wstęgi Bollingera

Wstęgi Bollingera (ang. Bollinger Bands) to wskaźnik rysujący wstęgi zgodnie ze strukturą ceny. Jego celem jest zdefiniowanie względnych definicji cen wysokich i niskich. Co oznacza, że gdy cena znajduje się w pobliżu górnej wstęgi jest wysoka i można spodziewać się jej spadku, natomiast gdy cena znajduje się w pobliżu dolnej wstęgi jej wartość jest niska i można spodziewać się jej wzrostów.



Rysunek 2.11 Bollinger Bands
Źródło: Bollinger, Bollinger on Bollinger Bands, McGraw-Hill New York 2001, s xxi

Podstawą wstęg Bollingera jest wskaźnik średniej kroczącej, stanowi on również środkową wstęgę opisywanego wskaźnika. Domyślna wartość średniej kroczącej użytej w wstęgach Bollingera to 20 okresów. Szerokość wstęg definiowana jest przez miarę zmienności określanej jako odchylenie standardowe. Dane do obliczenia zmienności to te same dane, które były użyte do obliczeń średniej kroczącej. Wstęga górna i dolna rysowane są w odległości dwóch odchyleń standardowych od średniej.

$$Wstęga\ górna = wstęga\ środkowa + 2 * odchylenie\ standardowe$$

$$Wstęga\ środkowa = 20 - okresowa\ średnia\ ruchoma$$

$$Wstęga\ dolna = wstęga\ środkowa - 2 * odchylenie\ standardowe$$

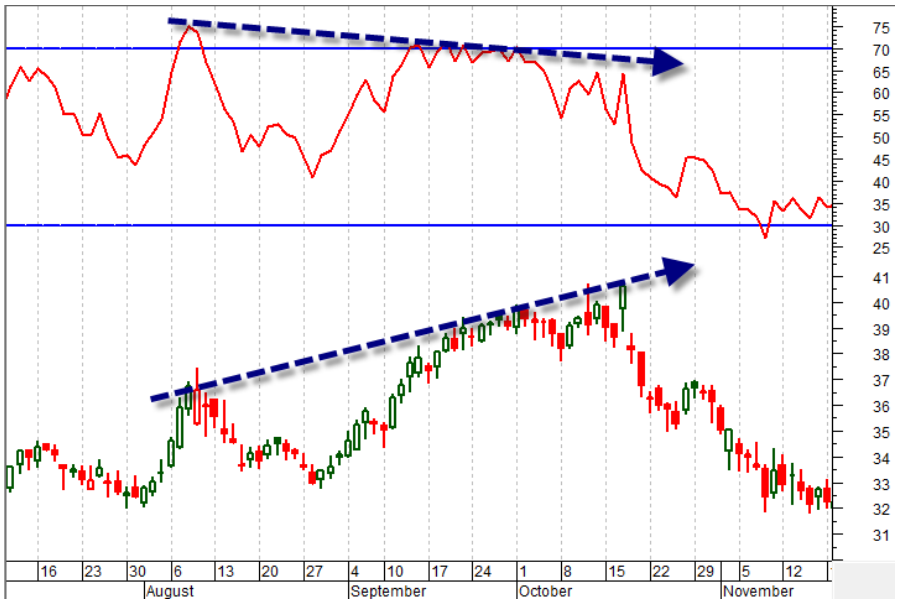
Analiza z pomocą wstęg Bollingera sprowadza się do odczytywania formacji występujących na rynku w okolicy wstęg, analizy przecięć ceny z wstęgami czy wybić ceny poza wstęgi [3].

2.4.3.3 Oscylatory

Szczególnej uwagi wymagają oscylatory. Wskaźniki należące do tej rodziny są pomocnicze przy analizie trendu, którego skuteczność jest największa, kiedy nie ma dużych ruchów cen. Konstrukcja wszystkich oscylatorów jest podobna, jest to rysowany na bieżąco wykres o osi na poziomie zero, która dzieli wykres na górną i dolną połowę lub o maksymalnej wartości 100 i minimalnej 0. Oscylatory sygnalizują możliwość zmiany kierunku ruchu ceny w momencie, gdy wartość oscylatora zbliża się do ustalonych wartości ekstremalnych. Gdy wartość oscylatora zbliża się maksimum mowa jest o wykupieniu rynku, natomiast gdy jego wartość zbliża się do minimum mówimy o wysprzedaniu rynku. Oscylatory posiadają poziomy pomagające określić wykupienie bądź wysprzedanie rynku [4].

Najczęściej poszukiwane na oscylatorach są przecięcia poziomów oraz dywergencje. Przecięcia poziomów w oscylatorze oznaczają wykupienie bądź wysprzedanie w zależności od przeciętego poziomu. Im bardziej przekroczony jest poziom tym silniejszy jest sygnał wykupienia bądź wysprzedania co zwiastuje zmianę kierunku ruchu ceny.

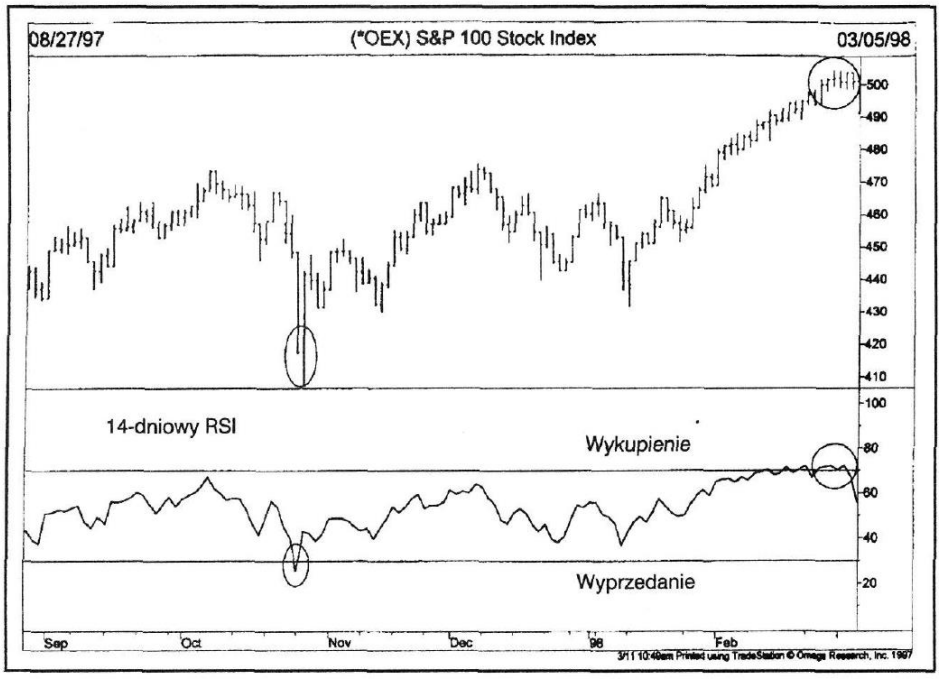
Dywergencje są to sytuacje, kiedy wskaźnik i wykres ceny podążają w przeciwnych kierunkach. Przykładowo, gdy cena rośnie, a wartość wskaźnika spada może być to interpretowane jako sygnał odwrócenia trendu. Dywergencja jest silniejsza im większa jest rozbieżność między wskaźnikiem a wykresem [4].



Rysunek 2.12 Dywergencja
Źródło: https://www.learningmarkets.com/how-to-trade-bullish-and-bearish-technical-divergences/ (dostęp 11.07.2017)

2.4.3.4 RSI – Relative Strength Index

Jest to jeden z najpopularniejszych wskaźników siły względnej należący do oscylatorów. Zakres wartości wskaźnika to od 0 do 100, stały zakres daje możliwość dokonywania porównań. Wadą tego wskaźnika jest to że podczas silnych trendów i gwałtownych zmian ceny jego skuteczność mocno spada. Zdefiniowane są w nim poziomy domyślne 70 i 30 przekroczenie, których uważane jest za sygnał wykupienia bądź wyprzedania rynku.



Rysunek 2.13 Wskaźnik RSI
Źródło: Murphy, Analiza techniczna rynków finansowych, WIG-PRESS Warszawa 1999, s 210

Do inicjalizacji wskaźnika podajemy okres w postaci cyfry, przykładowo 14 to okres domyślny oraz najpopularniejszy i oznacza on, że do obliczeń przyjmowany jest okres 14 dni. Do obliczania wartości wskaźnika używany jest następujący wzór:

$$RSI = 100 - \frac{100}{1 + RS} \tag{1}$$

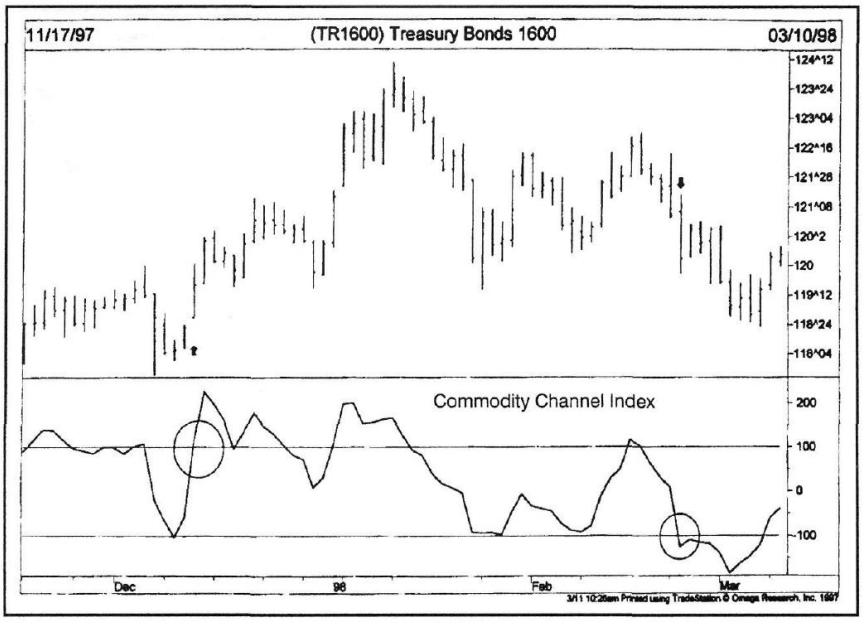
$$RS = \frac{Średnia\ wartość\ wzrostu\ cen\ zamknięcia\ z\ x\ dni}{Średnia\ wartość\ spadku\ cen\ zamkięcia\ z\ x\ dni} \tag{2}$$

Do otrzymania średniej wartości wzrostu cen zgodnie ze wzorem (2) trzeba dodać wszystkie punkty zyskane podczas wzrostów oraz podczas przyjętego okresu (w tym przypadku 14 dni), a następnie podzielić ich sumę przez przyjęty okres, czyli 14. Aby potrzymać średnią wartość spadku analogicznie należy zsumować wszystkie punkty straty podczas dni spadkowych i podzielić je przez 14. Siłą względna oznaczona jako RS obliczana jest poprzez podzielenie średniej wzrostu przez średnią spadku, a następnie ta wartość jest podstawiana pod wzór na obliczenie RSI (1).

Oscylator ten jest tym bardziej czuły i większa jest jego amplituda im jego okres jest krótszy. Dzięki temu do analizy krótkoterminowej można obniżyć okres by zwiększyć czułość oscylatora, natomiast aby uniknąć zakłóceń przy analizie długoterminowej można zwiększyć okres wskaźnika [4].

2.4.3.5 CCI - Commodity Channel Index

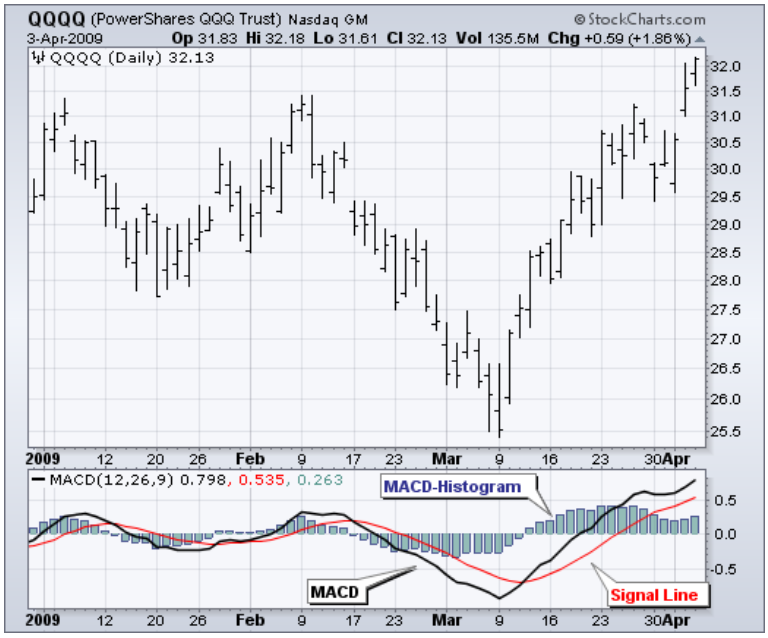
Jest to skonstruowany przez Donalda R. Lamberta wskaźnik stworzony do wykrywania początków i końców trendów na giełdach towarowych, ale jest stosowany również na innych. Porównuje on aktualną cenę średnią ze średnią dla podanego okresu, domyślnie jest to 20 dni. Po tym otrzymane wartości oscylatora są normalizowane za pomocą dzielnika bazującego na przeciętnym odchyleniu. Wynikiem tego jest wartość wskaźnika CCI oscylująca zazwyczaj od +100 do -100. Twórca wskaźnika zaleca przyjmowanie długich pozycji kupna gdy CCI wskazuje wartość powyżej +100 aż do momentu gdy wróci do zakresu, oraz krótkich sprzedaży gdy spada poniżej -100 [1].



Rysunek 2.14 Wskaźnik CCI
Źródło: Murphy, Analiza techniczna rynków finansowych, WIG-PRESS Warszawa 1999, s 208

2.4.3.6 MACD - Moving Average Convergence/Divergence

Wskaźnik wynalazł Gerald Appel pod koniec lat siedemdziesiątych. Jest to bardzo popularny wskaźnik, zamieniający dwie średnie ruchome (wskaźniki śledzące trend) w oscylator mierzący zmiany ceny. W rezultacie otrzymany został wskaźnik dający informację o kierunku trendu oraz zmienności ceny. MACD oscyluje powyżej i poniżej poziomu 0, wskazując dywergencje, konwergencje oraz przecięcia linii. MACD nie ma ograniczeń co do wartości, więc nie jest odpowiednim wskaźnikiem do oceny wykupienia wyprzedania rynku.



Rysunek 2.15 Wskaźnik MACD
Źródło: http://stockcharts.com/school/doku.php?id=chart_school:technical_indicators:moving_average_convergence_divergence_macd (dostęp na 17.07.2017)

Wskaźnik MACD składa się z linii MACD, która jest różnicą między dwoma wskaźnikami wykładniczej średniej kroczącej (ang. EMA – Exponential Moving Average), przykładowo od 12 dniowej jest odejmowania 26 dniowa. Linia MACD przedstawia konwergencję i dywergencję tych dwóch linii. Linia sygnałowa jest średnią linii MACD z 9 okresów, natomiast histogram przedstawia różnicę między linią MACD a linią sygnałową.

Dywergencja zachodzi, gdy linie średnich kroczących na podstawie których linia MACD jest obliczana się rozchodzą. Wartość linii MACD rośnie, kiedy 12 dniowa średnia jest większa od 26 dniowej i maleje, gdy sytuacja jest odwrotna. W obu przypadkach podczas dywergencji linia MACD oddala się od poziomu zero, natomiast podczas konwergencji oznaczającej sytuację, kiedy linie średnich kroczących podążają w tym samym kierunku linia MACD zbliża się do poziomu 0.

MACD daje sygnały w momencie przecięcia linii MACD z poziomem 0, gdy przecięcie następuje od góry jest to sygnał do sprzedaży, a gdy od dołu to do kupna. Przecięcia linii MACD z linią sygnałową w ekstremach również są sygnałami, które mogą zwiastować zmianę trendu. Dodatkowymi informacjami są dywergencje względem wykresu, gdy linie MACD i sygnałowa poruszają się w przeciwnym kierunku niż wykres [2].

2.5 Uczenie maszynowe

Uczenie maszynowe to dziedzina powiązana z statystyką, informatyką, logiką, teorią informacji, teorią prawdopodobieństwa i nie tylko. Jest częścią większego pojęcia jakim jest sztuczna inteligencja. Celem uczenia maszynowego jest automatyczna detekcja istotnych informacji z dużych źródeł danych. Automatyzacja niesie za sobą konieczność uczenia się tak aby dostarczane dane były jak najbardziej precyzyjne. Narzędzia uczenia maszynowego uczą się i adoptują w taki sposób by zwracać celne wyniki [9].

2.5.1 Pojęcie uczenia się w kontekście maszynowym

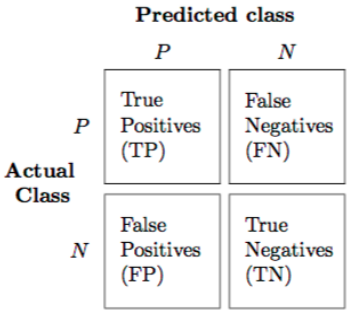
Proces uczenia się zawiera kilka ważnych elementów, które spełniać musi również system określany jako uczący się. Takimi elementami są zmiany o charakterze uczenia się, to oznacza, że system powinien wprowadzać zmiany w interpretacji danych w taki sposób by poprawić swoje działanie. Poprawa za pośrednictwem zmian powinna wynikać z działania systemu, czyli to on musi zdecydować, że dana zmiana wpłynie pozytywnie na jego działanie i ją zastosować. Nie mogą być to zmiany funkcjonalności systemu czy inne zmiany zewnętrzne. Takie zmiany wynikające z wnętrza systemu można określić jako doświadczenie [6].

W kontekście maszyn proces uczenia w dużym uproszczeniu polega na analizie zbioru danych treningowych na których algorytm uczy się, czyli wprowadza zmiany w interpretacji danych w celu poprawienia wyników. Dzieje się to za pomocą odpowiednio zbudowanego algorytmu, w którym zmiany wynikają z wnętrza systemu na podstawie analizy danych wprowadzonych z zewnątrz. Na koniec otrzymany jest wynik, czyli informacje na bazie wprowadzonego zbioru danych [9].

2.5.2 Podstawowe pojęcia

Poniżej zostaną opisane podstawowe pojęcia związane z uczeniem maszynowym użyte w pracy [17, 18]:

- Atrybuty – rodzaj wartości jaki przyjmuje określona kolumna w zbiorze danych. Atrybuty można określić jako symboliczne, czyli określona jest lista typów wartości jakie może przyjmować: numeryczny, tekstowy, format daty,
- Filtry – narzędzia modyfikujące w określony sposób zestawy danych, mogą m.in usuwać lub dodawać atrybuty, usuwać instancje, dyskretyzować,
- Instancja – reprezentuje pojedynczy wiersz w zestawie danych i składa się z określonej liczby atrybutów,
- Klasa – atrybut decyzyjny do którego przypisywana jest instancja po klasyfikacji,
- Klasyfikator – każdy algorytm uczenia maszynowego dostarczony przez wekę. Z użyciem dostarczonego zestawu danych tworzy model, na podstawie którego mogą być klasyfikowane nowe przykłady,
- Macierz pomyłek – macierz prezentująca wynik klasyfikacji, wiersze odpowiadają poprawnym klasom decyzyjnym, a kolumny decyzjom podjętym przez klasyfikator,



Rysunek 2.16 Schemat macierzy pomyłek
Źródło: https://rasbt.github.io/mlxtend/user_guide/evaluate/confusion_matrix/ (dostęp na 18.09.2017)

w zakres macierzy pomyłek wchodzą takie pojęcia jak:

  - True Positive (TP) – poprawne określenie klasy pozytywnej, klasa pozytywna określona jako pozytywna
  - False Positive (FP) – nieprawidłowe określenie klasy pozytywnej, klasa pozytywna określona jako negatywna
  - False Negative (FN) – nieprawidłowe określenie klasy negatywnej, klasa pozytywna określona jako negatywna

    - True Negative (TN) – poprawnie określenie klasy negatywnej, klasa negatywna określona jako negatywna

- Model – jest sposobem klasyfikacji danych wytrenowany przez klasyfikator. Po otrzymaniu na wejściu instancji, na wyjściu otrzymana zostanie informacja do jakiej klasy pasuje. Modele mogą być dynamiczne co oznacza, że mogą trenować się w miarę otrzymywania nowych danych,
- Precyzja – jest to wartość poprawnie sklasyfikowanych instancji (TP) w relacji do wszystkich danych sklasyfikowanych jako pozytywne (TP + FP), określona wzorem (3):

$$precision = \frac{TP}{TP + FP} \tag{3}$$

- Czułość (ang. Recall) – wartość poprawnie sklasyfikowanych instancji (TP) ze wszystkich instancji danej klasy (TP + FN), określona wzorem (4):

$$recall = \frac{TP}{TP + FN} \tag{4}$$

- Walidacja krzyżowa – metoda testowania dzieląca zestaw danych na określoną ilość podzbiorów losowych oraz dla każdego buduje model i go testuje,
- Zbiór treningowy – zbiór danych wykorzystywany przez klasyfikator do treningu modelu,
- Zbiór testowy – zbiór danych wykorzystywany przez klasyfikator do testowania modelu,
- Zestaw danych – wszystkie dane wprowadzone do weki, na podstawie których wykonywane są działania, składa się z instancji.

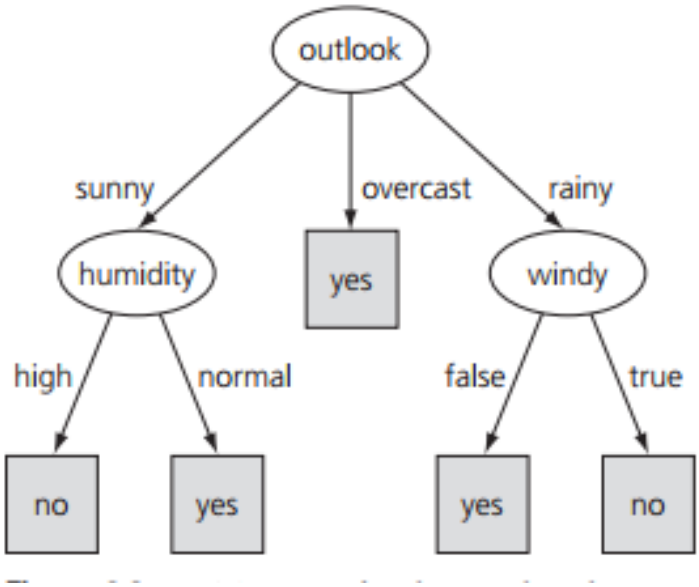
2.5.3 Przedstawienie badanych algorytmów

Podczas badań użyta została większość algorytmów uczenia maszynowego jakie dostarcza biblioteka Weka (opisana w 3.2). Poniżej zostaną przedstawione skrótowo algorytmy wybrane jako najlepsze.

2.5.3.1 Lasy losowe

Lasy losowe (ang. Random Forests) jest to metoda klasyfikacji z użyciem wielu algorytmów drzew decyzyjnych. Polega na tworzeniu wielu drzew decyzyjnych bazując na losowym zestawie danych. Każde drzewo decyzyjne klasyfikuje problem a następnie końcowa decyzja jest podejmowana w drodze głosowania większościowego na najbardziej popularna klasę.

Drzewo decyzyjne jest to skierowany graf acykliczny, oparty na strukturze drzewiastej. Każde drzewo posiada węzły, a może być nimi decyzja lub stan i gałęzie, które odpowiadają wariantom. Z węzła może wychodzić tyle gałęzi, ile jest wariantów decyzji, a każdy liść oznacza decyzję. Klasyfikacja rozpoczyna się od korzenia drzewa, następnie podąża po gałęziach a kończy na liściach, czyli ostatnim elemencie drzewa.



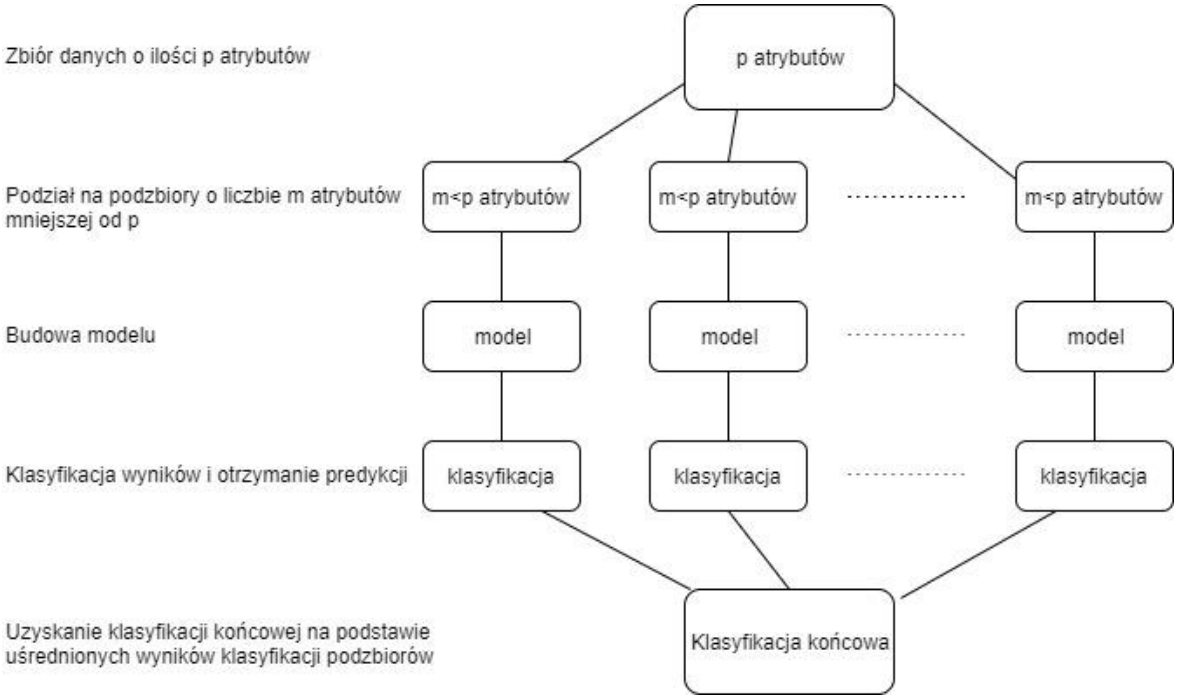
Rysunek 2.17 Przykładowe drzewo
Źródło: Witten,Eibe,Mark., Data Mining: Practical Machine Learning Tools and Techniques

Użytkownik ma możliwość zdefiniowania ilości drzew klasyfikujących. Tworzone są różne drzewa z użyciem różnych algorytmów, za pomocą algorytmu bagging, który generuje zróżnicowany zestaw klasyfikatorów użytych w lasach.

Zasada działania algorytmu bagging polega na wygenerowaniu określonej liczby klasyfikatorów ekspertów dla podzbioru zadań. Z całego zestawu danych jest losowany podzbiór poprzez losowanie ze zwracaniem a następnie dla niego jest tworzony ekspert, czyli drzewo, następnie losowany jest kolejny podzbiór i tworzony jest kolejne drzewo. Proces ten powtarzany jest określoną ilość razy, a na koniec stworzone drzewa są używane do głosowania. Decyzja posiadająca najwięcej głosów jest przyjmowana [13].

2.5.3.2 Metoda podprzestrzeni losowych

Metoda podprzestrzeni losowych (ang. Random Subspace Method, RSM) wykorzystywana przy danych z dużą ilością atrybutów, kiedy wyszukiwanie pośród takiej ilości zmiennych może być długie i trudne, aby zidentyfikować istotne zmienne.



Rysunek 2.18 Wykres etapów algorytmów RSM
Źródło: opracowanie własne

Metoda operuje na losowo generowanych podzbiorach atrybutów, które należą do całości atrybutów dostępnych w zestawie danych. Na danych o mniejszej ilości atrybutów algorytm uczący może poradzić sobie lepiej niż przy ich dużej ilości a tym samym uzyskać celniejsze informacje. Ze względu na dużą ilość atrybutów losowanie podzbiorów odbywa się do momentu, gdy wszystkie zostaną pokryte. W rezultacie otrzymujemy taką samą ilość podzbiorów, ile było losowań. Dla każdego podzbioru generowany jest model zgodnie z wybranym algorytmem. Każdy podzbiór jest klasyfikowany a wyniki predykcji wszystkich klasyfikatorów są sumowane i otrzymywana jest ich średnia [11].

3. Opis wykorzystanych narzędzi

Poniższy dział opisuje narzędzia jakie zostały wykorzystane do badań. Podrozdział 3.1 opisuje narzędzie w języku Java dedykowane do przeprowadzania analizy technicznej oraz budowania strategii. Kolejny podrozdział 3.2 opisuje narzędzie stanowiące zestaw algorytmów uczenia maszynowego, które zostało wykorzystane do budowy i testowania modelu.

3.1 Technical Analysis for Java (Ta4j)

Ta4j jest to biblioteka napisana w języku Java typu open source na licencji MIT, służąca do przeprowadzania analizy technicznej. Dostarcza również podstawowe komponenty do tworzenia, oceny oraz stosowania strategii inwestowania. Jest to biblioteka napisana w całości w języku Java działająca na wersjach 1.6 lub wyższych. Zawiera ponad 100 gotowych oraz przetestowanych wskaźników analizy technicznej, oraz silnik do tworzenia strategii inwestycyjnych.

3.1.1 Podstawy korzystania z TA4j

Aby używać biblioteki wystarczy dołączyć plik .jar do projektu. Podstawowymi obiektami w bibliotece są TimeSeries czyli obiekty zawierające zagregowane dane o określonych przedziałach czasowych. Każde zakończenie przedziału czasowego jest reprezentowane przez obiekt typu Tick. TimeSeries można porównać do wykresu giełdowego o określonym przedziale czasowym.

Obiekt Tick jest obiektem zawierającym informacje o pojedynczym zakończonym przedziale czasowym. Jest reprezentacją świecy japońskiej i zawiera informacje o cenie otwarcia, cenie zamknięcia, najwyższej i najniższej cenie oraz wolumenie.

```
DateTime endTime = DateTime.now();
List<Tick> ticks = Arrays.asList(
    new Tick(endTime, 105.42, 112.99, 104.01, 111.42, 1337),
    new Tick(endTime.plusDays(1), 111.43, 112.83, 107.77, 107.99, 1234),
    new Tick(endTime.plusDays(2), 107.90, 117.50, 107.90, 115.42, 4242),
    //...
);
TimeSeries series = new TimeSeries("my_2014_series", ticks);
```

Rysunek 3.1 Tworzenie obiektu TimeSeries
Źródło: https://github.com/mdeverdelhan/ta4j/wiki/Time%20series%20and%20ticks (dostęp na 24.07.2017)

Jednym ze sposobów utworzenia obiektu TimeSeries jest utworzenie listy obiektów Tick reprezentujących świece i na podstawie tej listy utworzenie obiektu TimeSeries za pomocą konstruktora.

Kolejnym ważnym elementem biblioteki są dostarczone wskaźniki analizy technicznej. Aby stworzyć wskaźnik zazwyczaj potrzeba obiektu TimeSeries dla wskaźników podstawowych, natomiast dla wskaźników bardziej złożonych wymagane są inne wskaźniki.

```
SMAIndicator sma = new SMAIndicator(closePrice, 20);
StandardDeviationIndicator sd = new StandardDeviationIndicator(closePrice, 20);
BollingerBandsMiddleIndicator bbm = new BollingerBandsMiddleIndicator(sma);
BollingerBandsUpperIndicator bbu = new BollingerBandsUpperIndicator(bbm, sd, Decimal.valueOf(2.5));
BollingerBandsLowerIndicator bbl = new BollingerBandsLowerIndicator(bbm, sd, Decimal.valueOf(2.5));

EMAIndicator ema = new EMAIndicator(closePrice, 9);
MACDIndicator macd = new MACDIndicator(ema, 13, 26);
```

Rysunek 3.2 Wskaźniki
Źródło: opracowanie własne

Dodatkowo konstruktory wskaźników przyjmują wartości liczbowe, które mogą oznaczać odchylenia czy okresy czasowe jakie przyjmuje wskaźnik. Wszystko zależnie od rodzaju wskaźnika. Wartości wskaźników są obliczane dynamicznie w momencie wywołania metody get (int index) na obiekcie wskaźnika.

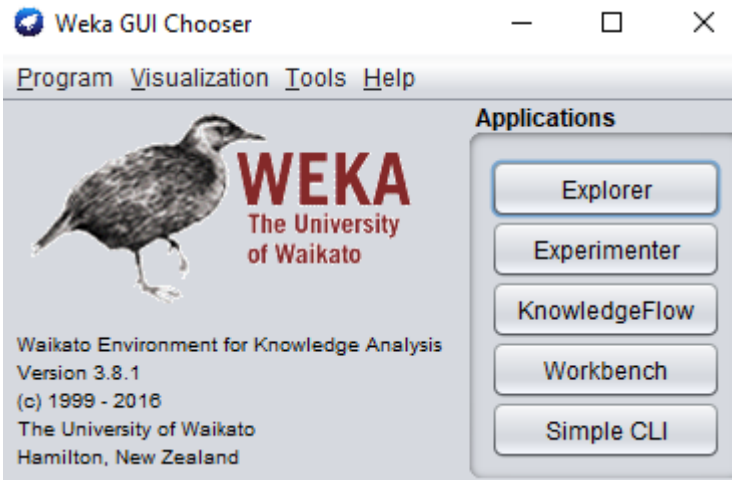
Innymi możliwościami biblioteki jest tworzenie strategii, czyli zestawu reguł do podejmowania decyzji kupna/sprzedaży oraz tester strategii za do sprawdzania poprawności stworzonych strategii z użyciem historycznych danych [12].

3.2 Weka

Weka jest to aplikacja dostarczająca wiele algorytmów uczenia maszynowego do eksploracji danych. Z weki można korzystać za pomocą przygotowanego interfejsu graficznego lub jako biblioteki Java, którą można załączyć do projektu i prosto z kodu używać zawartych algorytmów. Częścią weki są narzędzia do wstępnego przetwarzania danych, klasyfikacji, regresji, klastrowania, relacji oraz wizualizacji. Jest również przystosowana do rozwijania nowych schematów uczenia maszynowego. Rozwijana jest przez Uniwersytet Waikato w Nowej Zelandii jako projekt open source pod licencją GNU (GNU General Public License, GPL). Weka dostarcza również gotowe aplikacje, które wybiera się podczas uruchamiania programu, są to:

- Explorer - środowisko do badania danych z użyciem weki,

- Experimenter – środowisko do wykonywania eksperymentów oraz testów statystycznych pomiędzy schematami uczenia,
- KnowledgeFlow – ta sama funkcjonalność co Explorer dodatkowo obsługująca interfejs „drag and drop” oraz obsługująca uczenie inkrementacyjne,
- Workbench – aplikacja zawierająca wszystkie powyższe funkcjonalności
- SimpleCLI – wierz poleceń umożliwiający bezpośrednie wykonywanie komend weki, przeznaczony dla systemów niedostarczających własnego wiersza poleceń.



Rysunek 3.3 Weka – aplikacje
Źródło: opracowanie własne

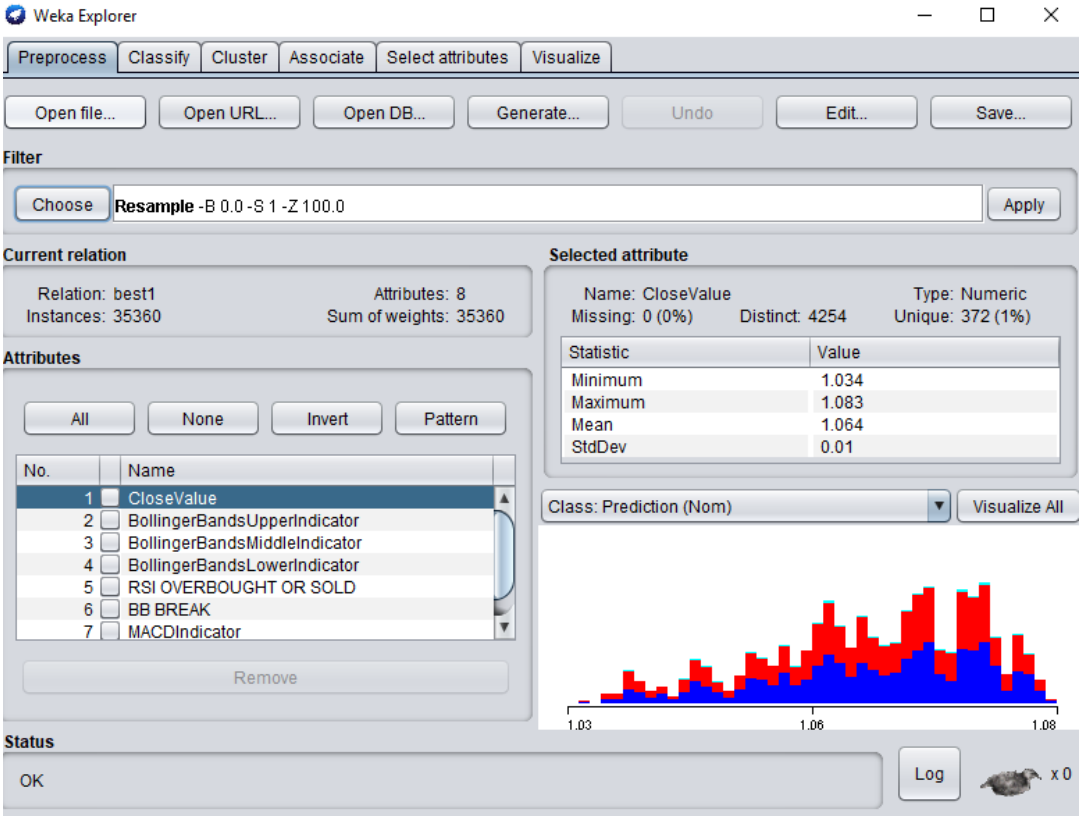
Funkcjonalności weki opisywane w kolejnych podrozdziałach będą na podstawie aplikacji Explorer, z uwagi na największe wykorzystanie właśnie tej aplikacji podczas badań [17].

3.2.1 Wstępne przetworzenie danych

Do wstępnego przetwarzania danych służy sekcja „Preproscess”, umożliwia ona prowadzenie danych w postaci różnych rozszerzeń plików takich jak „.arff” czy „.csv”, również dane mogą zostać wprowadzone z odnośnika URL, bazy danych lub wygenerowane za pomocą wbudowanego generatora. W sekcji wstępnego przetwarzania danych zawiera się kilka innych podsekcji [17]:

- **Filter** - daje możliwość filtrowania danych za pomocą dostarczanych przez wekę gotowych filtrów,
- **Current Relation** - prezentuje podstawowe dane wprowadzonego zbioru danych: nazwę, ilość atrybutów, ilość wierszy danych oraz sumę wag,
- **Attributes** – prezentuje atrybuty jakie posiada dostarczony zbiór danych oraz umożliwia usuwanie wybranych,

- **Selected attribute** – prezentuje parametry wybranego atrybutu, jego ekstrema, ilość różnych powtarzających oraz unikalnych wartości i dodatkowo graf reprezentujący zbieżność wartości z wartościami wybranego innego atrybutu.



Rysunek 3.4 Sekcja preprocess
Źródło: opracowanie własne

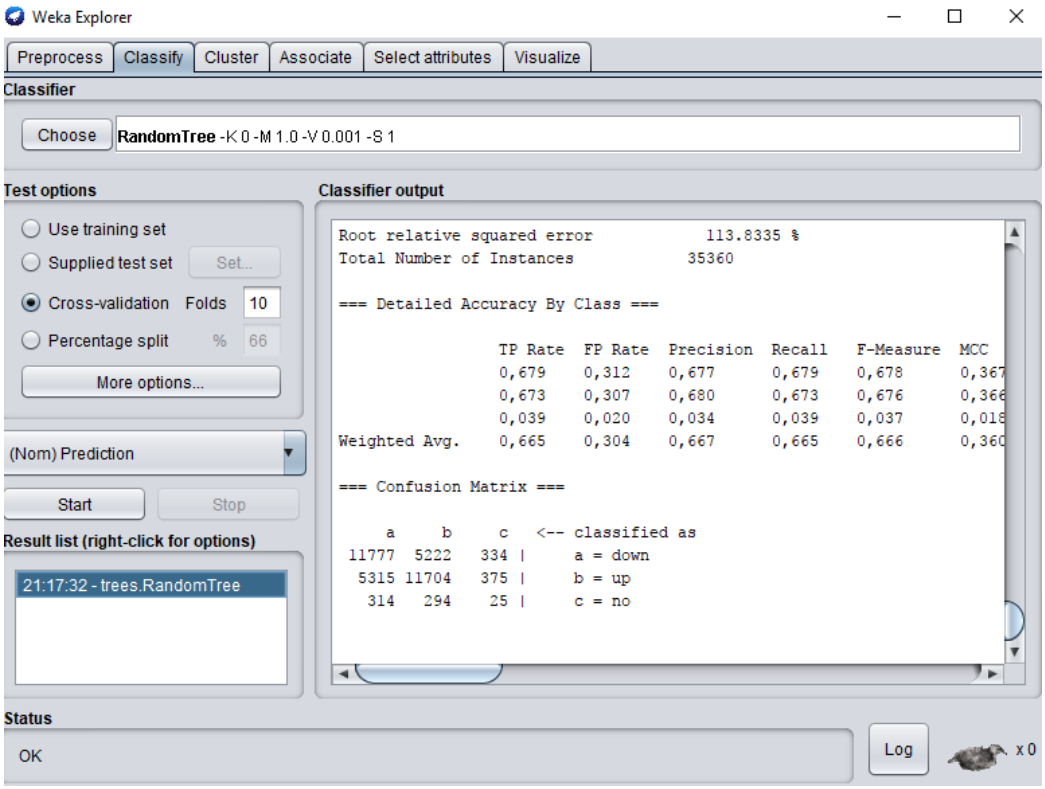
3.2.2 Klasyfikacja danych

Do klasyfikacji danych służy sekcja „Classify", która składa się z następujących podsekcji:

- **Classifier** – umożliwia wybór klasyfikatora oraz ustawienie jego parametrów po kliknięciu w nazwę,
- **Test options** – wybór opcji testowania stworzonego modelu. Jest to opcja pojedynczego wyboru, gdzie wybrane może zostać użycie tylko zbioru treningowego, użycie dodatkowego zbioru testowego, walidacja krzyżowa z możliwością określenia ilości zbiorów testowych na które zostaną podzielone dane treningowe oraz podział procentowy,
- **Result list** – lista wyników klasyfikacji, zapisywana po każdym wykonaniu dzięki czemu jest możliwość sprawdzenia poprzednich klasyfikacji kliknięcie prawym przyciskiem myszy na wynik wywoła listę czynności jakie można

wykonać z stworzonym modelem m.in zapisanie modelu, wizualizacja modelu, analiza kosztów,

- **Classifier output** – dokładny wynik klasyfikacji z m.in opisem modelu, macierzy pomyłek, dokładności względem klas.



Rysunek 3.5 Sekcja klasyfikacji danych
Źródło: opracowanie własne

3.2.3 Pozostałe funkcjonalności

Pozostałe funkcjonalności były używane w mniejszym stopniu do badań jednak wciąż warto o nich wspomnieć, są to sekcje [17]:

- **Cluster** – umożliwia wykonanie procesu klastrowania. Jest to proces grupowania obiektów w taki sposób by obiekty w tej samej grupie były jak najbardziej do siebie podobne. Interfejs jest niemal identyczny jak w sekcji „Classify" z tą różnicą, że zamiast wyboru algorytmów klasyfikujących są do wyboru algorytmy klastrujące.
- **Associate** – sekcja uczenia maszynowego bazującego na regułach. Polega na poszukiwaniu silnych relacji pomiędzy zmiennymi dostarczonymi w zbiorze danych. Interfejs umożliwia wybór algorytmu oraz przeglądanie listy rezultatów.

- **Select attributes** – jest to narzędzie do oceny wartości atrybutów, szczególnie przydatne, gdy zbiór danych posiada wiele atrybutów i niektóre mogą wprowadzać zamęt. Sekcja ta dostarcza narzędzia odpowiednie do selekcji wartościowych atrybutów. Dostarczona jest możliwość wybrania sposobu oceny atrybutów, metody wyszukiwania, sposobu selekcji atrybutów ze zbioru oraz okno z listą rezultatów i wynikiem selekcji atrybutów.
- **Visualize –** sekcja prezentuje matrycę wykresów rozproszenia dla relacji każdego atrybutu z każdym. Atrybuty są oznaczone kolorami względem klas do których należą. Wykresy mogą być powiększane, zarówno jak i obiekty atrybutów na wykresie.

4. Metodologia

Początkowo rozdział przedstawia charakter badań i ich zakres. Kolejny podrozdział opisuje ogólny proces badań przybliżający sposób ich przebiegu. Następnie kolejno omówione są użyte do pracy dane początkowe oraz proces analizy, któremu zostały one poddane, aby wydobyć z nich informacje. Omówiony jest wygenerowany zbiór danych użyty przy pracy z algorytmami uczenia maszynowego. Natomiast ostanie podrozdziały opisują kolejno etapy badań.

4.1 Cel i przedmiot badań

Praca magisterska ma na celu zbadanie rezultatów krótkoterminowej prognozy kursów giełdowych z wykorzystaniem algorytmów uczenia maszynowego. Do zakresu badań należy:

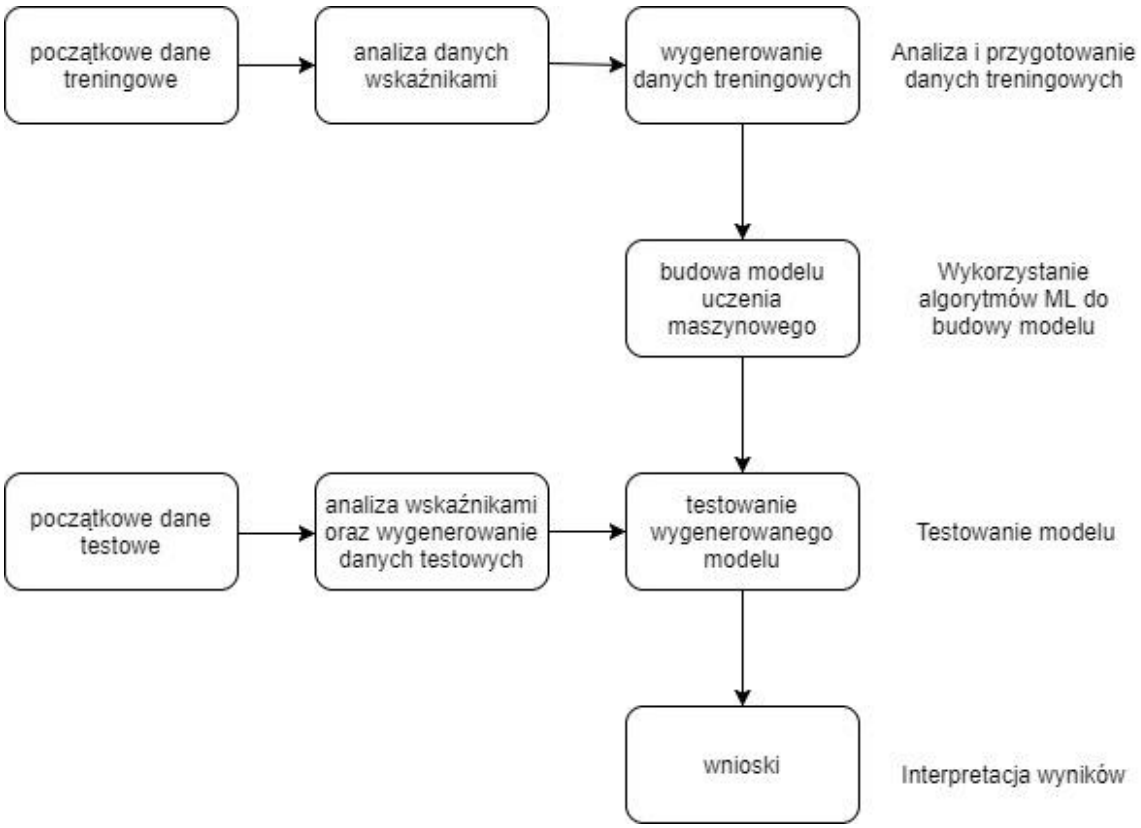
- poszukiwanie optymalnego algorytmu prognozy zmian wybranych par walutowych,
- symulację gry na opcjach binarnych z wykorzystaniem zbudowanego modelu.

Przedmiotem badań były wybrane algorytmy uczenia maszynowego dostarczone przez oprogramowanie Weka, spośród których został wybrany jeden użyty do symulacji gry na opcjach binarnych. Prognozowane były przyszłe kierunki ruchu ceny na giełdzie, czyli informacja czy cena z końcem określonego przedziału czasowego będzie wyższa czy niższa niż w czasie wykonywania prognozy. Przewidywania były wykonywane z pomocą algorytmów uczenia maszynowego na podstawie danych giełdowych przekształconych z użyciem dostępnych wskaźników analizy technicznej. Dane te służyły jako zbiory wejściowe do algorytmów klasyfikacyjnych.

Podstawą do podejmowania decyzji odnośnie skuteczności algorytmów były wyniki prognozy kierunku ruchu w górę, czyli precyzja prognozy ruchu w górę również określana jako „Precision UP”, czułość prognozy ruchu w górę określana również jako „Recall Up”. Oraz wyniki prognozy kierunku ruchu w dół, czyli precyzja prognozy ruchu w dół określana również jako „Precision Down” i czułość prognozy ruchu w dół określana również jako „Recall Down”.

4.2 Ogólny przebieg badań

Przebieg badań składał się z pewnych etapów następujących po sobie. Dane wykorzystane do badań były początkowo w stanie wymagającym obróbki i interpretacji co zostało wykonane przez program napisany w języku Java. Wygenerowany nowy zestaw danych bazujący na zbiorze początkowym. Służący jako zbiory treningowe i testowe dla algorytmów uczenia maszynowego, będącymi przedmiotem badań w celu określenia ich skuteczności przy przewidywaniu przyszłych ruchów cen rynków walutowych.



Rysunek 4.1 Etapy pracy
Źródło: Opracowanie własne

Tak jak na powyższym rysunku etapów pracy (rys. 4.1) całość można podzielić na cztery etapy. Pierwszy to przygotowanie danych treningowych, gdzie dostarczone do napisanego programu Java początkowe dane treningowe są mapowane do odpowiednich obiektów a następnie wykonywane na nich są obliczenia wskaźników analizy technicznej, które częściowo stanowią dane treningowe. Na podstawie wskaźników generowane są sygnały uzupełniające dane treningowe. Całość jest zbierana do jednego pliku csv. Kolejnym etapem jest budowa modelu, gdzie wcześniej wygenerowany plik csv jest ładowany do programu weka. Tam za pomocą filtrów modyfikowany jest zbiór treningowy oraz generowany jest

model uczenia maszynowego. Następny etap to testowanie modelu, które odbywa się zaraz po jego wygenerowaniu. Początkowy zbiór danych testowych przechodzi identyczną drogę analizy i generowania sygnałów. Otrzymany gotowy zbiór danych testowych jest wskazywany podczas ustawiania procesu generowania modelu za pomocą wybranego algorytmu uczenia maszynowego. Na postawie testowania otrzymywana jest macierz pomyłek oraz inne szczegółowe informacje o wygenerowanym modelu gotowe do analizy w ostatnim kroku badań.

4.3 Omówienie użytych danych

Do badań użyte zostały dane historyczne rynków walutowych. Ze wszystkich dostępnych par walutowych wybrana została para EUR/USD ze względu na jej największą popularność i dużą zmienność, co się wiąże z największą ilością transakcji dokonywanych właśnie na tej parze walutowej. Dostępne są dane od roku 2000 jednak ze względu na charakter badań, gdzie celem jest krótkoterminowa prognoza kierunku zmiany kursu walut wykorzystane zostały dane nowe pochodzące z stycznia roku 2017. Zestaw danych pochodzący z jednego miesiąca składa się z dokładnie takiej samej liczby plików, ile jest tygodni w danym miesiącu. Poniższe dane (rys. 4.2) jest to mała część danych pochodzących z pierwszego tygodnia stycznia 2017r. Pojedynczy plik posiada w tym przypadku ponad 353 tyś. wierszy, podobna liczba wierszy występuje w pozostałych plikach. Ilość wierszy zależy od ilości ruchów ceny w tygodniu, ponieważ każdy wiersz odpowiada pojedynczej zmianie w cenie pary walutowej.

C353783

	A
1	lTid,cDealable,CurrencyPair,RateDateTime,RateBid,RateAsk
2	5562501259,D,EUR/USD,2017-01-02 02:00:15.287000000,1.051550,1.051810
3	5562501274,D,EUR/USD,2017-01-02 02:00:17.037000000,1.051870,1.052070
4	5562501289,D,EUR/USD,2017-01-02 02:00:17.287000000,1.051880,1.052250
5	5562501307,D,EUR/USD,2017-01-02 02:00:18.287000000,1.051850,1.052200
6	5562501312,D,EUR/USD,2017-01-02 02:00:18.537000000,1.051870,1.052220
7	5562501341,D,EUR/USD,2017-01-02 02:00:24.287000000,1.051890,1.052220
8	5562501344,D,EUR/USD,2017-01-02 02:00:24.537000000,1.051860,1.052220
9	5562501350,D,EUR/USD,2017-01-02 02:00:25.037000000,1.051890,1.052220
353772	5578613827,D,EUR/USD,2017-01-06 16:59:30.380000000,1.053080,1.053580
353773	5578613854,D,EUR/USD,2017-01-06 16:59:32.380000000,1.053060,1.053560
353774	5578613859,D,EUR/USD,2017-01-06 16:59:32.630000000,1.053070,1.053570
353775	5578613882,D,EUR/USD,2017-01-06 16:59:33.880000000,1.053000,1.053500
353776	5578613889,D,EUR/USD,2017-01-06 16:59:34.630000000,1.053070,1.053570
353777	5578613894,D,EUR/USD,2017-01-06 16:59:34.880000000,1.053060,1.053560
353778	5578613906,D,EUR/USD,2017-01-06 16:59:35.130000000,1.053070,1.053570
353779	5578613917,D,EUR/USD,2017-01-06 16:59:35.380000000,1.053080,1.053580
353780	5578613933,D,EUR/USD,2017-01-06 16:59:36.630000000,1.053060,1.053560
353781	5578613949,D,EUR/USD,2017-01-06 16:59:36.880000000,1.053070,1.053570
353782	5578613961,D,EUR/USD,2017-01-06 16:59:37.630000000,1.053060,1.053560

Rysunek 4.2 Przykład danych
Źródło: opracowanie własne

Dane dostarczone są w formacie CSV, a szczegóły jakie przedstawia każdy wiersz to identyfikator, para walutowa, data zmiany ceny, cena kupna i cena sprzedaży. Podczas pracy z danymi nie wszystkie były istotne, wykorzystane zostały poszczególne kolumny odpowiadające za datę zmiany ceny, rodzaj pary walutowej oraz cena kupna. Jakość danych jest wysoka ze względu na to, że zanotowana jest każda minimalna zmiana ceny. Nieraz jest to kilka wierszy dotyczących jednej sekundy.

4.4 Omówienie procesu obróbki danych do sygnałów

Aby uzyskać odpowiednie dane do badań został napisany program w języku Java. Dane wejściowe wymagają specjalnej konwersji by pasowały do formatu użytego w bibliotece do analizy technicznej. Podrozdział przedstawia w jaki sposób aplikacja z danych początkowych tworzy świece japońskie, jak oblicza wartości wskaźników na podstawie świec oraz w jaki sposób na podstawie wskaźników są generuje sygnały.

4.4.1 Tworzenie świec japońskich

Wyżej przedstawione dane wejściowe, zanim zostaną zmienione na świece japońskie muszą zostać najpierw czytane do listy obiektów StockRecord, który reprezentuje pojedynczy wiersz z otrzymanego pliku CSV. Następnie dla każdego takiego wiersza wykonywana jest akcja przyporządkowania do jakiej świecy należy dany wiersz. Tworzony

Jeżeli wiersz należy do okresu **to:**

Pobierz i porównaj jego parametry z parametrami świecy, gdy przekracza ekstrema świecy podmień wartość.

W przeciwnym razie:

Zapisz aktualną świecę, stwórz nową i przepisz wartości wiersza do świecy

Koniec

Rysunek 4.3 Pseudokod algorytmu klasyfikującego wiersz do świecy
Źródło: opracowanie własne

jest obiekt MutableTick reprezentujący pojedynczą świecę z informacją zakresu przedziału czasowego jaki obejmuje. Po czym każdy wiersz z danych jest przyporządkowywany do świecy, a jego wartości nadpisują odpowiednie ekstrema świecy, jeżeli je przekraczają. Gdy

pojawi się wiersz danych należący do kolejnego przedziału czasowego, aktualna świeca jest zapisywana i tworzona jest następna do której są zapisywane wiersza należące do kolejnego okresu.

4.4.2 Generowanie sygnałów

Na podstawie otrzymanego zestawu świec japońskich tworzone są obiekty wskaźników, które dla każdej świecy obliczają odpowiednią wartość liczbową. Obliczone wartości już same w sobie są istotnymi informacjami do oceny przyszłego kierunku ceny, niemniej jednak na podstawie niektórych wskaźników generowane są dodatkowo sygnały kupna i sprzedaży.

4.4.2.1 Sygnał RSI

Wskaźnik RSI może generować sygnał prawdopodobnego spadku ceny reprezentowany przez liczbę jeden oraz wzrostu ceny reprezentowany przez liczbę minus jeden. Gdy wartość sygnału wynosi zero oznacza to informację o braku okazji do decyzji. Zakładając, że wartość poziomu górnego wskaźnika RSI jest oznaczony jako $\boldsymbol{x}$ a poziom dolny jako $\boldsymbol{y}$ i wartość wskaźnika RSI to $\boldsymbol{rsi}$ to generowany sygnał jest opisany następującymi równaniami (5):

$$\begin{cases} 1 = rsi > x \\ 0 = x < rsi < y \\ -1 = rsi < y \end{cases} \tag{5}$$

Gdy wartość wskaźnika rsi jest wyższa niż poziom górny oznacza to wykupienie rynku i generowany jest sygnał spadku ceny oznaczony jako jeden, natomiast w odwrotnej sytuacji, gdy wartość wskaźnika rsi jest poniżej poziomu dolnego oznacza to wysprzedanie rynku i generowany jest sygnał wzrostu ceny, czyli minus jeden. Natomiast gdy wskaźnik porusza się pomiędzy poziomami generowany jest sygnał zero będący informacją o braku okazji do decyzji.

Dodatkowo możliwe było wygenerowanie skalowalnego sygnału RSI. Wymagane jest podanie dla wskaźnika parametru oznaczającego zwiększenie progu. Na jego podstawie będzie on generował sygnał zwiększając lub zmniejszając wartość liczbową wraz z przekroczeniem kolejnych progów określonych przez pięciostopniową skalę. Podanie parametru zwiększenia progu o wartości 0.2 powoduje generowanie sygnałów o wyższej wartości po przekroczeniu kolejnego progu zwiększonego bądź zmniejszonego o dodatkowo 20% bazowej wartości określonego ekstremum w wskaźniku.

4.4.2.2 Sygnał CCI

Kolejny wskaźnik generujący sygnały to CCI w którym generowanie sygnału działa podobnie jak w powyższym RSI. Możliwe jest wygenerowanie trzech sygnałów. Sygnał oznaczony jako jeden oznacza sygnał do spadku ceny, minus jeden to sygnał wzrostu ceny zero to brak sygnału.

$$\begin{cases} 1 = cci > x \\ 0 = x < cci < y \\ -1 = cci < y \end{cases} \tag{6}$$

W CCI również występują poziomy górny i dolny na podstawie których generowane są sygnały. Gdy wartość wskaźnika oznaczona jako ***cci*** przekracza poziom górny oznaczony jako ***x*** generowana jest jedynka, czyli sygnał spadku ceny. W odwrotnej sytuacji, gdy ***cci*** przekracza poziom dolny oznaczony jako ***y*** oznacza to sygnał wzrostu ceny, czyli minus jeden. Gdy wartość wskaźnika porusza się między poziomami generowane jest zero jako brak sygnału.

Podobnie jak w RSI możliwe było wygenerowanie skalowalnego sygnału CCI, który po podaniu parametru zwiększenia progu dla wskaźnika generował będzie sygnał zwiększając lub zmniejszając jego wartość wraz z przekroczeniem kolejnych progów określonych przez pięciostopniową skalę.

4.4.2.3 Sygnał Wstęg Bollgingera

Na podstawie wskaźnika wstęg Bollingera również generowane są sygnały. Reprezentacja podobna jest jak w poprzednich. Generowane są trzy sygnały: jeden, minus jeden, zero. Wskaźnik składa się z linii górnej i dolnej, które śledzą ruch ceny. Gdy wartość ceny przekroczy wartość wskaźnika wstęgi górnej generowany jest sygnał jeden, czyli sygnał spadku ceny. W przeciwnym wypadku, gdy wartość ceny przekroczy dolną wstęgę generowany jest sygnał minus jeden sygnalizujący wzrost ceny. Gdy cena porusza się pomiędzy wstęgami generowane jest 0 jako brak sygnału. Metodę generowania sygnału można opisać następującym układem równań, gdzie ***p*** wyraża wartość ceny, ***bbu*** wartość górnej wstęgi Bollingera a ***bbd*** wartość dolnej wstęgi Bollingera.

$$\begin{cases} 1 = p > bbu \\ 0 = bbu < p < bbd \\ -1 = p < bbd \end{cases} \tag{7}$$

4.4.3 Wygenerowane dane

Dane wygenerowane przez aplikację stworzoną na potrzeby badań składają się z wielu elementów takich jak cechy świecy, wartości wskaźników i sygnały. Z miesięcznego zestawu ruchów ceny, który zawierał ponad 350 tysięcy wierszy, po zamianie na świece zostało ponad 35 tysięcy wierszy, które stanowią dane wyjściowe aplikacji generującej. Będą służyły jako zbiór wejściowy dla algorytmu uczenia maszynowego.

Dane wyjściowe jakie otrzymywane są po analizach to:

- Candle body - wartość ciała świecy,
- Upper shadow - wartość cienia górnego,
- Lower shadow - wartość cienia dolnego,
- Is Bullish – czy świeca jest wzrostowa, jest to jednoczesna informacja czy świeca jest spadkowa tj. świeca wzrostowa nie może być spadkową,
- BB Break Scaled – przebicie wstęgi Bollingera skalowane w zależności od wielkości świecy ponad wstęgą. Skala reprezentuje wartości od -5 do 5,

Candle body	Upper shadow	Lower shadow	is Bullish	BB BREAK SCALED
0.000170	0.000180	0.000170	0	0
0.000010	0.000000	0.000000	-1	0
0.000030	0.000000	0.000000	-1	0
0.000010	0.000010	0.000040	0	0
0.000010	0.000010	0.000010	0	0
0.000000	0.000000	0.000010	-1	0
0.000080	0.000000	0.000000	-1	0
0.000080	0.000010	0.000020	-1	0

Rysunek 4.4 Wygenerowane dane 1/3
Źródło: opracowanie własne

- RSI – wartość wskaźnika RSI,
- RSI extreme touched – informacja czy RSI osiągnęło ekstremum, jeżeli tak to wartość 1 lub -1 w zależności od osiągniętego ekstremum, w innym przypadku 0,
- RSI extreme scaled – informacja czy RSI osiągnęło ekstremum wartość reprezentowana jest w skali od -5 do 5 w zależności od przebicia poziomu,

- CCI – wartość wskaźnika CCI,
- CCI channel break – informacja czy CCI przebiło kanał, jeżeli tak to wartość 1 lub -1 w zależności od osiągniętego ekstremum, w innym przypadku 0,
- CCI channel break scaled – informacja czy CCI przebiło kanał, wartość reprezentowana jest w skali od -5 do 5 w zależności od przebicia poziomu,
- MACD Indicator – wartość wskaźnika MACD,

RSI	RSI SCALED	RSI EXTREME TOUCHED	RSI EXTREME SCALED	CCI	CCI SCALED	CCI CHANNEL BREAK	CCI CHANNEL BREAK SCALED	MACDIndicator
4	4	0	0	0	0	0	0	-0.000001295039874411302982731554 1
4	4	0	0	1	1	0	0	1.37828804664723032069970 9E-7
4	4	0	0	0	0	0	0	0.000003219299415465778147993891 5
3	3	0	0	0	0	0	0	0.000007546565792848218004402927 4
3	3	0	0	0	0	0	0	0.000012780905226522962371120876 6
3	3	0	0	0	0	0	0	0.000018472874478026484753477758 2
3	3	0	0	0	0	0	0	0.000022840835332337201754323577 9
3	3	0	0	0	0	0	0	0.000025579947236096239527336428 7

Rysunek 4.5 Wygenerowane dane 2 /3
Źródło: opracowanie własne

- bbSizeBand – szerokość wstęgi Bollingera, czyli różnica pomiędzy wstęgą górną a dolną
- bbPircePosition – przeskalowania pozycja ceny w zakresie wstęg Bollingera.
- Prediction – faktyczne zachowanie ceny w okresie następującym. Wartość, która ma być przewidywana przez algorytm uczenia maszynowego.

bbSizeBand	bbPricePosition	Prediction
0.001189450671279933296120220887 0	28.455281638315428192148471989998	down
0.001162180392755841727573176691 6	23.846801999961212789905783951899	up
0.001132470259604590650040489841 0	19.544295442302679533139221636196	no
0.001106192618568303332252100723 6	20.453038304741670384872746967280	up
0.001082896283058350732743727040 6	21.402299509958198147070644061063	down
0.001060722005447000380624206217 2	20.635839340077290891475727807611	down
0.001032954238983409684911676795 8	5.400962751178231338282248120958 5	down
0.001006801867300612416756058764 6	-0.397297634213234328465637430968 38	down

Rysunek 4.6 Wygenerowane dane 2/3
Źródło: opracowanie własne

4.5 Omówienie procesu badań

Proces badań składał się z czterech etapów:

- wyznaczania klasyfikatorów na podstawie całego zbioru atrybutów by wybrać najskuteczniejsze
- wyznaczania atrybutów, gdzie za pomocą algorytmów oceniających były wyznaczane i odfiltrowane atrybuty najbardziej znaczące dla badań,
- następnie klasyfikacja za pomocą wybranych metod klasyfikujących
- testowanie modelu.

Dodatkowo w procesie badań można wyróżnić dwie części składające się z tych samych czterech etapów. Pierwsza część, gdy każdy etap był powtarzany dla tego samego zbioru danych tak by na domyślnych wartościach wyznaczyć zbiór atrybutów najbardziej znaczących, które będą stanowiły bazę do badań w kolejnym etapie i wyznaczyć klasyfikatory najlepiej radzące sobie z problemem tak by wyniki badań były najlepsze. W drugiej części zaczęto wprowadzać zmiany w ustawieniach wskaźników, sygnałów oraz parametrach klasyfikatorów tak by poprawić wcześniejsze rezultaty.

4.5.1 Wyznaczenie i filtrowanie atrybutów

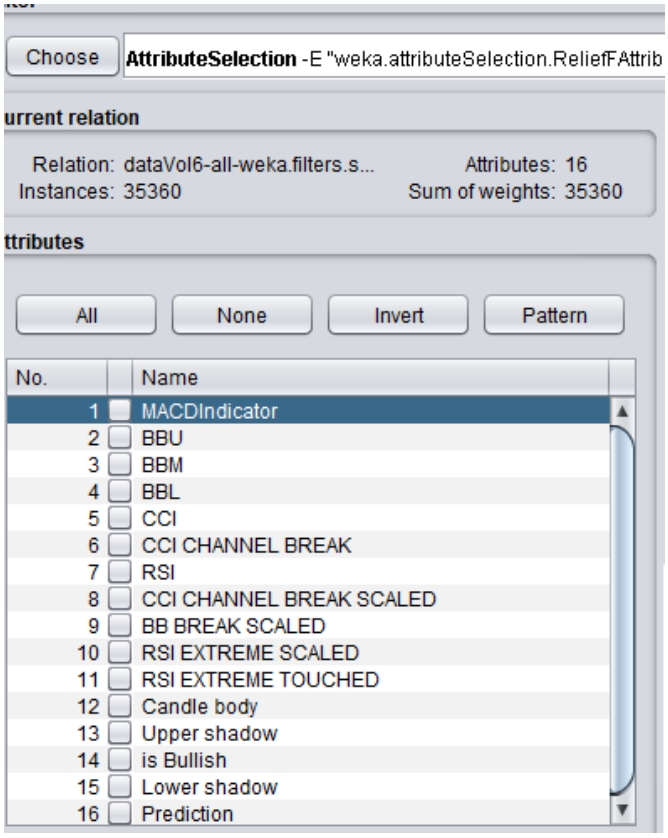
Etap wyznaczania atrybutów polegał na wykorzystaniu do wprowadzonych danych algorytmów selekcji atrybutów. Użyte zostało narzędzie „Select attributes” programu Weka, które dostarcza algorytmy umożliwiające selekcję. Zastosowany algorytm na zbiorze danych zwracał jako wynik listę potencjalnie najbardziej znaczących atrybutów bądź posegregowaną listę atrybutów względem oceny wykonanej przez algorytm. Selekcja odbywała się z wykorzystaniem wszystkich kombinacji metod oceniających i wyszukujących.

```
Attribute Subset Evaluator (supervised, Class (nominal): 16 Prediction)
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,3,4,7,8,9,10,12,15 : 9
                     Upper shadow
                     Lower shadow
                     is Bullish
                     BBL
                     BB BREAK SCALED
                     RSI
                     RSI EXTREME TOUCHED
                     CCI
                     MACDIndicator
```

Rysunek 4.7 Przykładowy wynik selekcji atrybutów
Źródło: opracowanie własne

Dostarczane wyniki przez algorytmy miały różną skuteczność przy klasyfikacji zbioru danych, dlatego do najlepszych wyników algorytmów traktowanych jako podstawa dobierane były dodatkowe atrybuty tak by zwiększyć trafność klasyfikacji w kolejnych etapach. Na podstawie dostarczonych wyników filtrowane były atrybuty używane do klasyfikacji w kolejnym etapie.



Rysunek 4.8 Zbiór atrybutów po użyciu filtra „AttributeSelection”
Źródło: opracowanie własne

Filtracja polegała na usunięciu atrybutów posiadających mniejszą wagę a pozostawienie tych posiadających większą wagę lub nałożeniu filtru „AttributeSelection”, który wykonuje wybrany algorytm selekcji atrybutów i automatycznie nanosi zmiany na listę atrybutów zgodnie z wynikiem algorytmu. Gdy algorytm zwraca podzbiór atrybutów jako wynik algorytmu, niezawarte w podzbiorze atrybuty były usuwane. Natomiast gdy wynikiem algorytmu była posortowana lista według oceny, kolejność była nanoszona na zbiór atrybutów.

4.5.2 Klasyfikacja

Klasyfikacja polegała na wykorzystaniu przefiltrowanego zestawu danych w poprzednim kroku do budowy modelu za pomocą algorytmów klasyfikujących. Początkowo do klasyfikacji używane były wszystkie algorytmy, którym na wejście był podawany ten sam zestaw danych po to, aby wyłonić te najbardziej skutecznie. Następnie klasyfikacja polegała na budowaniu modelu bazując na zmodyfikowanych danych z użyciem klasyfikatorów oznaczonych jako te o największej skuteczności dodatkowo zmieniając ustawienia parametrów klasyfikatora by zmaksymalizować trafność wyników.

```
=== Summary ===

Correctly Classified Instances        6923               57.5861 %
Incorrectly Classified Instances      5099               42.4139 %
Kappa statistic                          0.1666
Mean absolute error                      0.3325
Root mean squared error                  0.4062
Relative absolute error                 96.4032 %
Root relative squared error             97.8329 %
Total Number of Instances            12022

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  
                 0,614    0,443    0,571      0,614    0,592
                 0,559    0,390    0,581      0,559    0,570
                 0,000    0,000    0,000      0,000    0,000
Weighted Avg.    0,576    0,409    0,566      0,576    0,570

=== Confusion Matrix ===

    a    b    c   <-- classified as
 3615 2272    0 |    a = up
 2614 3308    0 |    b = down
  103  110    0 |    c = no
```

Rysunek 4.9 Przykładowy wynik klasyfikacji danych
Źródło: opracowanie własne

Klasyfikacja odbywała się z pomocą metod dostarczonych przez wekę czyli na podstawie danych treningowych, metodą cross-validation oraz metoda podziału zbioru danych w stosunku dwa do jednego czyli dwie trzecie instancji traktowanych było jak zbiór treningowy a pozostałe jako zbiór testowy. W tym etapie celem było uzyskanie modelu o najwyższych parametrach „precision" i „recall" w odniesieniu do wszystkich atrybutów decyzyjnych.

4.5.3 Testowanie

Etap testowania polegał na weryfikacji skuteczności modelu otrzymanego z etapu klasyfikacji. Początkowo etap testowania i klasyfikacji były połączone ze względu na dużą zmienność modeli podczas poszukiwania najbardziej trafnych. W miarę rozwoju prac etap testowania dotyczył mniejszej ilości modeli dzięki czemu możliwe było dokładniejsze testowanie na podstawie większej ilości zbiorów danych. Do weryfikacji skuteczności modelu używane były różnorodne zbiory testowe składające się z danych zarówno z krótkich okresów czasu takich jak tydzień oraz jak i z dłuższych nawet do dwóch miesięcy.

5. Rezultat badań

Rozdział piąty przedstawia wyniki badań, w kolejności w jakiej były uzyskiwane. Podrozdział 5.1 opisuje wyniki pierwszego etapu badań, czyli wyboru klasyfikatora. Kolejny podrozdział 5.2 przedstawia rezultaty poszukiwań najskuteczniejszych atrybutów. Podrozdział 5.3 prezentuje ustawienia wskaźników jakie zostały użyte by uzyskać najwyższą skuteczność modelu. Kolejne podrozdziały opisują wyniki symulacji uzyskanego modelu na danych z trzech miesięcy przy predykcji ruchów na pięć minut, trzydzieści minut oraz godzinę w przyszłość.

5.1 Klasyfikator

Badania wymagały zmniejszenia liczby klasyfikatorów, dla których będą przeprowadzane kolejne etapy pracy. Dlatego wybrane zostały konkretne klasyfikatory o największej skuteczności. Dla początkowego zestawu danych z nieprzefiltrowanymi atrybutami zostały wykonywane klasyfikacje z pomocą wszystkich klasyfikatorów jakie dostarcza weka. Wyniki zostały spisane w tabeli:

Algorytm	Precision up	Recall up	Precision down	Recall down	Precision no	Recall no
bayes						
BayesNet	0,494	0,642	0,524	0,401	0	0
NaiveBayes	0,501	0,509	0,519	0,538	0	0
NaiveBayes Multinominal	0	0	0,497	1	0	0
NaiveBayes Upadeable	0,501	0,509	0,519	0,538	0	0
functions						
Logistic	0,501	0,423	0,512	0,616	0	0
Multilayerpercepton	0,485	0,854	0,538	0,174	0	0
Simplelogistic	0,508	0,432	0,518	0,62	0	0
SMO	0,505	0,494	0,521	0,559	0	0
lazy						
Ibk	0,462	0,488	0,485	0,511	0,028	0,033
Kstar	0,489	0,632	0,518	0,394	0,016	0,004
LWL	0,494	0,694	0,531	0,354	0	0
meta						
AdaBoostM1	0,497	0,729	0,542	0,328	0	0
AttributeSelectedClassifier	0,493	0,6	0,519	0,438	0	0
Bagging	0,485	0,484	0,506	0,533	0	0
ClasifiactionViaRegression	0,508	0,432	0,518	0,619	0	0
VCPParameterSelecltion			0,497	1	0	0

FilteredClassifier	0,502	0,501	0,52	0,547	0	0
IterativeClassifierOtimizer	0,509	0,403	0,514	0,644	0	0
LogicBoost	0,509	0,403	0,514	0,644	0	0
MultiClassClassifier	0,501	0,423	0,511	0,615	0	0
MultiClassClassifierUpadeable	0,513	0,283	0,508	0,754	0	0
MultiScheme	0	0	0,497	1	0	0
RandomComitee	0,49	0,574	0,512	0,451	0	0
RandomizableFilteredClassifer	0,479	0,488	0,499	0,495	0	0
RandomSubSpace	0,503	0,481	0,519	0,534	0	0
Stacking	0	0	0,497	1	0	0
Vote	0	0	0,497	1	0	0
WeightedInstancesHandlerWrapper	0	0	0,497	1	0	0
misc						
InputMappedClassifier	0	0	0,497	1	0	0
rules						
DecisionTable	0,501	0,465	0,515	0,578	0	0
Jrip	0,518	0,373	0,517	0,683	0	0
OnerR	0,472	0,463	0,492	0,528	0	0
PART	0,491	0,609	0,519	0,422	0	0
ZeroR	0	0	0,497	1	0	0
trees						
DecisionStump	0,497	0,729	0,542	0,328	0	0
HoeffidingTree	0,49	0,742	0,53	0,296	0	0
J48	0,488	0,474	0,508	0,546	0	0
LMT	0,508	0,432	0,518	0,62	0	0
RandomForest	0,497	0,498	0,515	0,54	0	0
RandomTree	0,485	0,474	0,503	0,513	0,031	0,031
ReptTree	0,5	0,518	0,517	0,527	0	0

Tabela 5.1 Wyniki klasyfikacji na podstawowym zbiorze danych
Źródło: opracowanie własne

Na podstawie powyższej tabeli i selekcji atrybutów (opisane w kolejnym podrozdziale) wybrany jako najlepszy został klasyfikator lasów losowych (ang. Random Forest). Ten klasyfikator zwrócił najlepsze wyniki dla standardowego zbioru danych po procesie selekcji atrybutów. Algorytm lasów losowych nie był jedynym algorytmem o obiecujących wynikach. Na tym etapie badań, inne również miały wysoką precyzję w przewidywaniu kierunku ruchu jednak dodatkowo ważnym elementem, na który również zwracana była uwaga jest parametr „Recall" określający czułość algorytmu. Gdzie duża część algorytmów nie wypadła już tak dobrze. Szczególną uwagę należało jednak zwrócić na takie algorytmy jak Random Comittee, Random Subspace, Naive Bayes, Logistic, Simple Logistic, ClasifiactionViaRegression. IterativeClassifierOtimizer, Logic Boost, Decision Table, JRip oraz LMT. Algorytmy te miały wysoką skuteczność i prezentowały się początkowo lepiej

od lasów losowych. Szczególnie LMT, który wydawał się prezentować dużo lepiej niż wybrany algorytm. Jednak w dalszej części eksperymentów po selekcji atrybutów i modyfikacjach parametrów klasyfikatorów wyżej wymienione nie reagowały tak dobrze na wprowadzone zmiany jak lasy. Dlatego algorytm lasów losowych został wybrany jako główny na którym będą przeprowadzane dalsze eksperymenty.

5.2 Atrybuty

Początkowy zestaw atrybutów nie dawał dobrych rezultatów przy zastosowaniu żadnego z algorytmów. Zbyt duża ilość informacji dostarczanych do klasyfikatora powodowała jego błędne wskazania. Dane dostarczane do procesu klasyfikacji musiały przejść proces selekcji atrybutów dzięki któremu określone zostały te dające najlepsze rezultaty. Ograniczenie ilości atrybutów i eliminacja cech mających negatywny wpływ spowodowała wzrost skuteczności wybranego algorytmu.

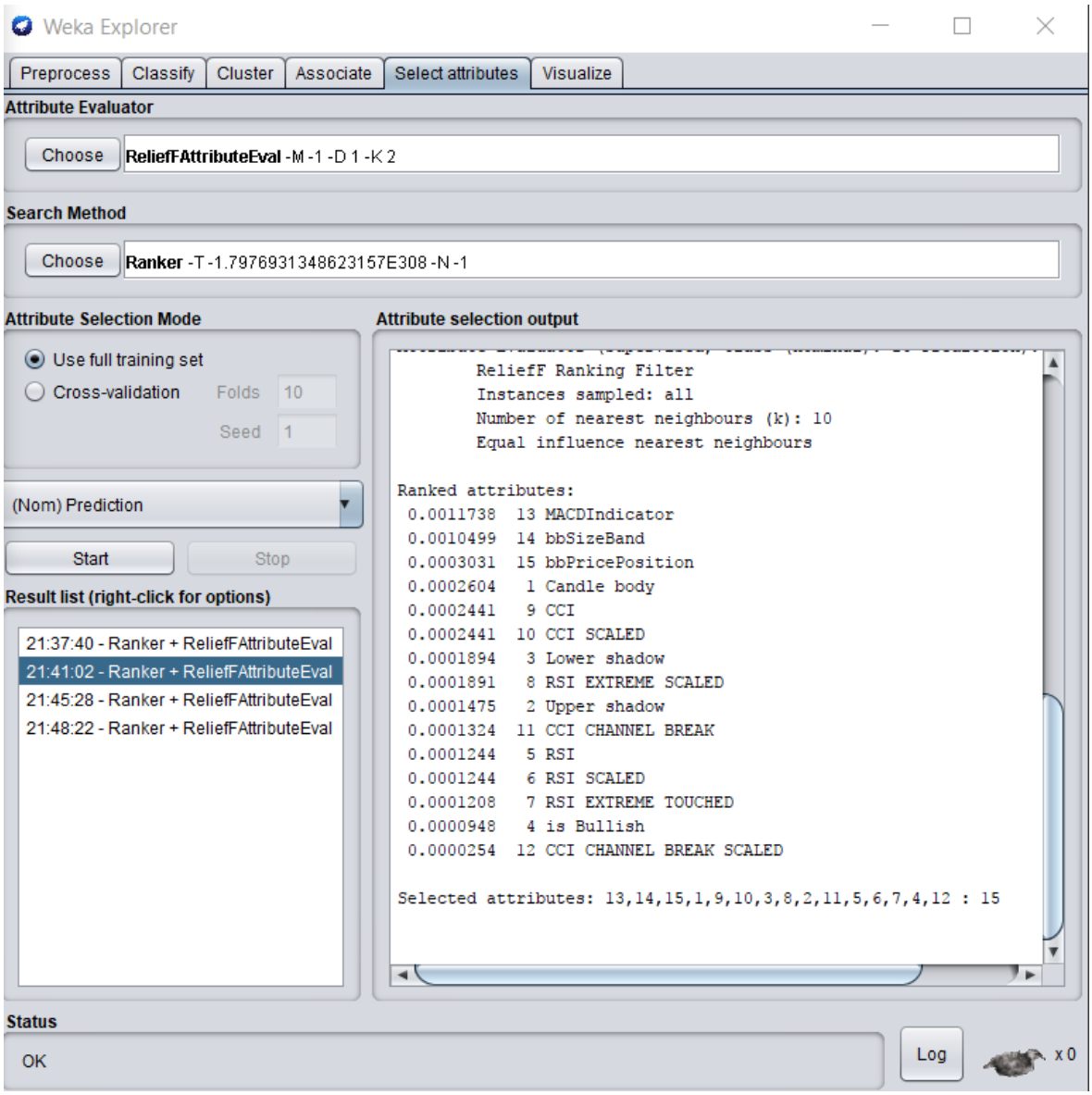
```
                 TP Rate   FP Rate   Precision   Recall
                 0,530     0,454     0,516       0,530
                 0,547     0,472     0,534       0,547
                 0,000     0,001     0,000       0,000
Weighted Avg.    0,525     0,451     0,511       0,525

=== Confusion Matrix ===

    a    b    c   <-- classified as
 2475 2189    2 |    a = up
 2200 2659    3 |    b = down
  121  134    0 |    c = no
```

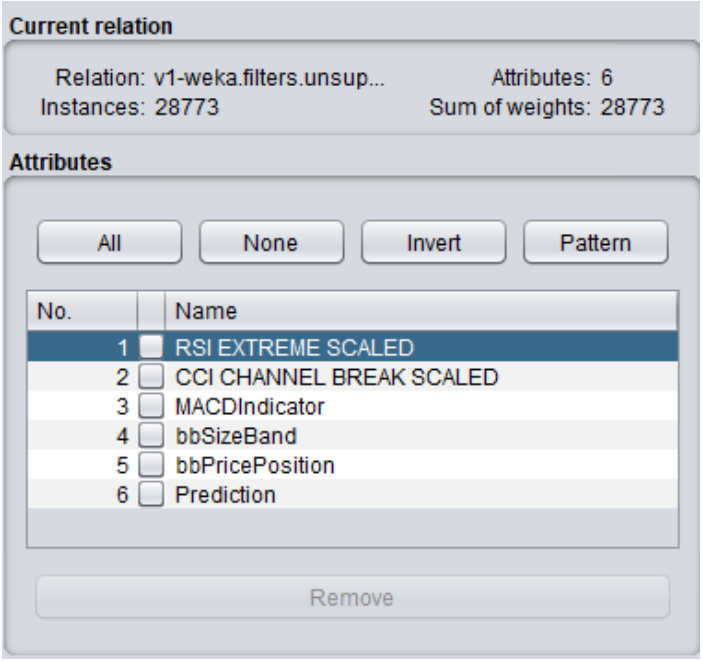
Rysunek 5.1 Skuteczność drzew losowych po selekcji atrybutów
Źródło: Opracowanie własne

Użyte do selekcji atrybutów narzędzie „Select attributes” z wykorzystaniem algorytmu selekcji cech „ReliefFAttributeEval” przy ustawieniu selekcji na podstawie 5 najbliższych sąsiadów z metodą wyszukiwania „Ranker” dały najtrafniejszą ocenę atrybutów.



Rysunek 5.2 Wynik selekcji atrybutów
Źródło: opracowanie własne

Na podstawie powyższej listy został określony finalny podzbiór atrybutów dający najlepsze rezultaty. Elementy zajmujące pierwsze miejsca w liście faktycznie okazały się tymi o największej skuteczności. Do określenia ostatecznej listy atrybutów wykonana została seria klasyfikacji oceniających skuteczność wybranych podzbiorów. Pomimo sugestii algorytmu selekcji atrybutów, że cechy o numerach 1,9,10,3 są skuteczniejsze niż atrybut o nr 8, czyli „RSI EXTREME SCALED" właśnie ten zwracał lepsze wyniki. Algorytm ocenił atrybut o numerze 11 „CCI CHANNEL BREAK" wyżej niż 12 „CCI CHANNEL BREAK SCALED" jednak do finalnego zbioru wybrana została druga cecha ze względu na większą możliwość modyfikacji jej ustawień i lepszy wynik w testach manualnych przy klasyfikacjach.



Rysunek 5.3 Ostateczny podzbiór atrybutów
Źródło: opracowanie własne

Finalnie został wybrany podzbiór sześciu następujących atrybutów: RSI EXTREME SCALED, CCI CHANNEL BREAK SCALED, MACDIndicator, bbSizeBand, bbPricePosition oraz Prediction jako atrybut decyzyjny pozostał bez zmian. Pozostałe atrybuty zostały odrzucone ze względu na ich negatywny wpływ na wyniki klasyfikacji.

5.3 Ustawienia wskaźników

Do badań określone zostać musiały również najlepsze ustawienia wskaźników, które zostały użyte do generowania danych. Zostało to wykonane po wyborze klasyfikatora i selekcji atrybutów, ponieważ badanie wpływu zmian atrybutów na wyniki klasyfikacji było bardzo czasochłonnym procesem. Również ilość klasyfikatorów, dla których badany był wpływ parametrów musiała być ograniczona do minimum, oraz atrybuty wykorzystywane do klasyfikacji musiały być sprecyzowane. Poniżej przedstawiona została skrócona tabela wyników modyfikacji wskaźników i ich wpływu na wyniki predykcji. Tabela przedstawia jedynie wyniki przewyższające najlepszy wynik predykcji po wyborze klasyfikatora i selekcji atrybutów.

bbu deviation val	bbl deviation val	RSI period	RSI up extreme	RSI down extreme	CCIi period	CCI up extreme	CCI down extreme	MACD	MACD EMA	precision up	recall up	precision down	recall down	precision no	recall no
2,5	2,5	14	70	30	20	150	-150	13/26	9	0,514	0,531	0,531	0,541	0	0
3	3	14	70	30	20	150	-150	13/26	9	0,519	0,529	0,535	0,553	0	0
2,8	2,2	14	70	30	20	150	-150	13/26	9	0,52	0,535	0,536	0,55	0,2	0
2,8	2,1	14	70	30	20	150	-150	13/26	9	0,519	0,534	0,537	0,55	0	0
2,8	2,2	14	70	30	20	150	-150	13/26	9	0,516	0,52	0,531	0,554	0,143	0
2,8	2,2	14	70	30	20	150	-150	13/26	9	0,518	0,533	0,536	0,548	0	0
2,8	2,2	10	70	30	20	150	-150	13/26	9	0,519	0,524	0,535	0,557	0	0
2,8	2,2	19	70	30	20	150	-150	13/26	9	0,518	0,529	0,534	0,551	0	0
2,8	2,2	14	65	30	20	150	-150	13/26	9	0,518	0,529	0,536	0,552	0	0
2,8	2,2	14	70	30	21	150	-150	13/26	9	0,516	0,528	0,533	0,548	0	0
2,8	2,2	14	70	30	20	150	-150	16/29	9	0,521	0,534	0,540	0,554	0	0
2,8	2,2	14	70	30	20	150	-150	17/30	9	0,521	0,533	0,541	0,556	0	0
2,8	2,2	14	70	30	20	150	-150	16/30	9	0,526	0,535	0,543	0,562	0,182	0,008
2,8	2,2	14	70	30	20	150	-150	17/28	9	0,529	0,541	0,546	0,561	0	0
2,8	2,2	14	70	30	20	150	-150	17/29	9	0,521	0,533	0,541	0,556	0	0
2,8	2,2	14	70	30	20	150	-150	16/30	7	0,522	0,537	0,542	0,555	0,111	0,004
2,8	2,2	14	70	30	20	150	-150	16/30	12	0,526	0,541	0,544	0,557	0,167	0,004
2,8	2,2	14	70	30	20	150	-150	16/30	13	0,526	0,541	0,547	0,559	0,167	0,004
2,8	2,2	14	70	30	20	150	-150	16/30	14	0,526	0,550	0,545	0,549	0,125	0,004
2,8	2,2	14	70	30	20	150	-150	16/30	15	0,531	0,562	0,552	0,549	0	0
2,8	**2,2**	**14**	**70**	**30**	**20**	**150**	**-150**	**16/30**	**16**	**0,532**	**0,561**	**0,554**	**0,553**	**0,167**	**0,004**
2,8	2,2	14	70	30	20	150	-150	16/30	17	0,532	0,545	0,550	0,565	0	0
2,8	2,2	14	70	30	20	150	-150	16/30	18	0,527	0,555	0,55	0,551	0	0
2,8	2,2	14	70	30	20	150	-150	16/30	19	0,529	0,548	0,549	0,558	0	0

Tabela 5.2 Wyniki wpływu zmian parametrów wskaźników
Źródło: opracowanie własne

W tabeli został wytłuszczony najlepszy wynik modyfikacji parametrów wskaźników. Poszczególne zmiany wartości parametrów nie wpływały znacznie na zmiany predykcji. Jednak większość zmian obniżała trafność algorytmu niż go podwyższała. Po wykonaniu niemal stu wariantów z użyciem różnych wartości parametrów wybrany został oznaczony zestaw. Najlepszy model osiągnął w przybliżeniu:

- Precyzję na poziome 53,2% określenia kierunku ruchu w górę, przy czułości wynoszącej 56,1%,
- Precyzję na poziome 55,4% określenia kierunku ruchu w dół, przy czułości wynoszącej 55,3%,
- Precyzję na poziome 16,4% określenia braku ruchu ceny, przy czułości wynoszącej 0,4%.

Model zbudowany został na następujących wartościach parametrów wskaźników:

- Bollinger Bands – wartość wskaźnika prostej średniej kroczącej wyrażająca długość okresu to 20, który użyty został do budowy linii wskaźnika BB. Wartość wskaźnika odchylenia standardowego wyrażająca długość okresu użytego do budowania górnej i dolnej linii to 20, oraz mnożnik odchylenia standardowego ustawiony został na 2,8 dla wstęgi górnej i 2,2 dla wstęgi dolnej,
- RSI – długość okresu ustawiona na wskaźniku wynosiła 14 dni. Natomiast wartość górnego ekstremum, po którym wyznaczane były sygnały kupna ustawiona została na 70, a dolnego do wyznaczania sygnałów sprzedaży na 30. Wartości generowanych sygnałów zwiększały się, gdy przekraczany był kolejny próg extremum zwiększony o 20% bazowego ekstremum.
- CCI - długość okresu ustawiona na wskaźniku wynosiła 20 dni. Natomiast wartość górnego ekstremum, po którym wyznaczane były sygnały kupna ustawiona została na 150, a dolnego do wyznaczania sygnałów sprzedaży na 150 poniżej zera
- MACD – wartość wskaźnika wykładniczej średniej kroczącej użytego do budowy MACD wyrażająca długość okresu ustawiona została na 16. Natomiast wartości dwóch średnich kroczących wymaganych przez wskaźnik to 16 i 30.

5.4 Testy

Poniżej zostaną opisane wyniki testów wykonanych z użyciem wcześniej uzyskanych parametrów badań tj. klasyfikatora, atrybutów i parametrów wskaźników. Testy zostały wykonane dla prognozy o pięć minut, o trzydzieści minut oraz o godzinę do przodu. Przyjęty okres danych do testów to trzy miesiące. Zostały użyte dane w całości bez podziału oraz z podziałem na miesiące, tygodnie i dni. W pojedynczym dniu wykonywane było wiele prognoz na podstawie których obliczana była dzienna precyzja.

Każdy rozdział prezentuje wynik modelu zbudowanego do prognozy, a następnie wyniki testów wykonanych na nim. Dla danych bez podziału oraz dla podziałów na miesiące i tygodnie wyniki zostały zaprezentowane w postaci tabeli ze względu na małą ilość wierszy. Natomiast dla podziału na dni wynik został zaprezentowany w postaci wykresu ze względu dużą ilość wierszy w tabeli. Przy omawianiu wyników dla podziału na tygodnie i dni ze względu na ich dużą ilość opisywane będą za pomocą zakresów.

5.4.1 Test - prognoza o pięć minut.

Poniżej na rysunku widać wynik uzyskanego modelu bazującego na parametrach opisanych w poprzednich podrozdziałach, który został użyty do przeprowadzenia symulacji przewidywania o pięć minut na podstawie danych testowych z trzech miesięcy.

```
=== Summary ===

Correctly Classified Instances        5310               54.2778 %
Incorrectly Classified Instances      4473               45.7222 %
Kappa statistic                          0.1094
Mean absolute error                      0.336
Root mean squared error                  0.4234
Relative absolute error                 96.1784 %
Root relative squared error            101.2312 %
Total Number of Instances             9783

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,561     0,450     0,532       0,561   0,546      0,112    0,573     0,541     up
               0,553     0,440     0,554       0,553   0,554      0,113    0,572     0,560     down
               0,004     0,001     0,167       0,004   0,008      0,022    0,556     0,033     no
Weighted Avg.  0,543     0,433     0,533       0,543   0,536      0,110    0,572     0,537

=== Confusion Matrix ===

    a    b    c   <-- classified as
 2619 2044    3 |    a = up
 2170 2690    2 |    b = down
  131  123    1 |    c = no
```

Rysunek 5.4 Podsumowanie wytrenowanego modelu do predykcji o 5 min
Źródło: opracowanie własne

Precyzja uzyskanego modelu wynosi 53.2% przy przewidywaniu ruchów w górę i 55.4% przy ruchach w dół. Czułość podjętych decyzji to 56.1% przy ruchach w górę i 55.3% przy ruchach w dół. Sklasyfikowane zostało poprawnie 54.28% instancji.

Niżej zostaną przedstawione wyniki skuteczności modelu w przewidywaniu kierunku ruchu giełdowego na pięć minut w przyszłość.

```
=== Summary ===

Correctly Classified Instances       45695               48.9098 %
Incorrectly Classified Instances     47732               51.0902 %
Kappa statistic                          0.0072
Mean absolute error                      0.3507
Root mean squared error                  0.4391
Total Number of Instances            93427

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,490    0,484    0,494      0,490    0,492      0,006    0,506     0,496     up
               0,517    0,508    0,485      0,517    0,501      0,008    0,506     0,485     down
               0,001    0,001    0,026      0,001    0,001      -0,000   0,515     0,030     no
Weighted Avg.  0,489    0,482    0,477      0,489    0,482      0,007    0,506     0,478

=== Confusion Matrix ===

     a     b     c   <-- classified as
 22461 23327    39 |     a = up
 21693 23232    35 |     b = down
  1329  1309     2 |     c = no
```

Rysunek 5.5 Podsumowanie ogólne symulacji 3-miesięcznej próby modelu predykcji o 5 min
Źródło: opracowanie własne

Jak widać na powyższym rysunku ogólna precyzja modelu nie odbiega znacznie od tej uzyskanej podczas treningu. Jest nieznacznie niższa co zdarza się często, że w testach wyniki nieznacznie się pogarszają. Sklasyfikowane zostało poprawnie 49% dostarczonych instancji. Precyzja ruchów w górę (49.4%) jest nieznacznie wyższa od precyzji ruchów w dół (48.5%). Precyzja i czułość predykcji braku ruchu jest bardzo mała ze względu na to, że brak ruchu na giełdzie zdarza się bardzo rzadko i przewidzenie takiego zachowania jest trudne.

miesiąc	Precision UP	Recall UP	Precision DOWN	Recall DOWN	Precision NO	Recall NO
marzec	0,494	0,482	0,498	0,535	0,04	0,001
kwiecień	0,492	0,487	0,479	0,513	0	0
maj	0,494	0,499	0,48	0,502	0,042	0,001

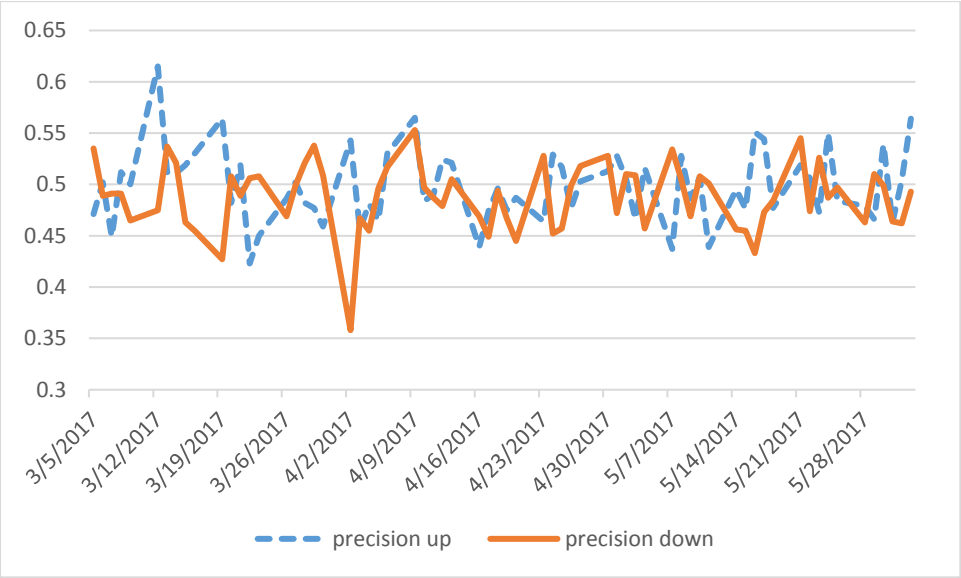
Tabela 5.3 Rezultaty symulacji predykcji o 5 min z podziałem na miesiące
Źródło: opracowanie własne

Na podstawie powyższej tabeli z wynikami z podziałem na każdy miesiąc można zauważyć, że model zachowuje niemal identyczną precyzję i czułość dla każdego z testowanych miesięcy. Widać nieznaczne różnice w precyzji predykcji kierunku w dół oraz w czułości przewidywań.

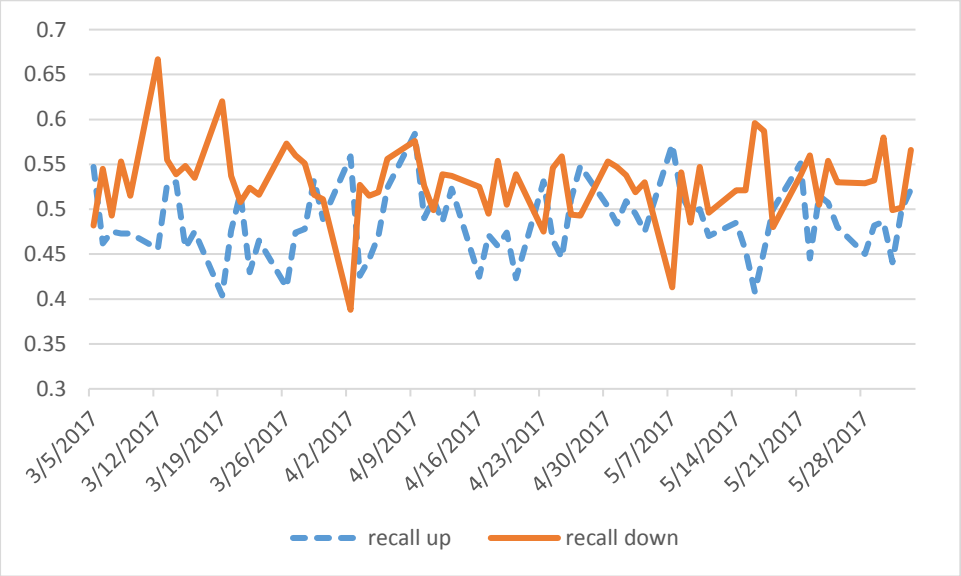
tydzień	Precision UP	Recall UP	Precision DOWN	Recall DOWN	Precision NO	Recall NO
1-marzec	0,492	0,477	0,49	0,526	0	0
2-marzec	0,523	0,493	0,494	0,552	0,25	0,01
3-marzec	0,481	0,467	0,491	0,527	0	0
4- marzec	0,481	0,492	0,518	0,535	0	0
1-maj	0,486	0,48	0,486	0,52	0,143	0,005
2-maj	0,501	0,509	0,494	0,521	0	0
3-maj	0,482	0,458	0,464	0,52	0,2	0,004
4-maj	0,504	0,488	0,477	0,52	0,25	0,005
1-czerwiec	0,498	0,484	0,486	0,529	0	0
2-czerwiec	0,491	0,497	0,498	0,52	0,125	0,005
3-czerwiec	0,518	0,456	0,457	0,546	0	0
4-czerwiec	0,494	0,481	0,496	0,533	0	0
5-czerwiec	0,505	0,485	0,487	0,536	0	0

Tabela 5.4 Rezultaty symulacji predykcji o 5 min z podziałem na tygodnie
Źródło: opracowanie własne

Tabela 5.4 przedstawia rezultaty z podziałem na tygodnie w miesiącu. Nadal są to uogólnione wyniki, ale można już zauważyć ich większe zróżnicowanie. Precyzja ruchów w górę wacha się od 48% do 52%, a precyzja ruchów w dół od 46% do 52% gdzie tylko dwa wyniki były poniżej 48%. Podobne wahania można zauważyć w czułości określania instancji. Dla ruchów w górę czułość mieści się w zakresie od 46% do 51%, natomiast dla ruchów w dół od 52% do 55%.



Rysunek 5.6 Wykres wyników precyzji predykcji o 5 min z podziałem na dni
Źródło: opracowanie własne



Rysunek 5.7 Wykres wyników czułości predykcji o 5 min z podziałem na dni
Źródło: opracowanie własne

Najdokładniej zróżnicowanie wyników widać na dwóch powyższych wykresach przedstawiających podział ze względu na dni. Zgodnie z prezentowanymi wykresami precyzja ruchów w górę i w dół mieści się w zakresie od 45% do 55% z pojedynczymi wynikami wychodzącymi poza zakres. Podobnie wyglądają wyniki czułości, gdzie zakres dla ruchów w gorę wynosi od 42% do 55% oraz dla ruchów w dół 47% do 59%. Tak jak w wynikach z podziałem na tygodnie i tutaj widać większą czułość w ocenie ruchów w dół. Anomalie w postaci pojedynczych wzrostów i spadków predykcji oraz czułości nie są niczym niepokojącym. Ze względu na dużą zmienność rynku w momencie wydarzeń

politycznych wykresy stają się nieprzewidywalne dla wskaźników co wpływa na wyniki przy silnych ruchach. Natomiast w momentach konsolidacji rynku i braku silnych trendów wskaźniki zyskują na skuteczności.

5.4.2 Test - prognoza trzydzieści minut

Poniżej na rysunku widać wynik uzyskanego modelu bazującego na parametrach opisanych w poprzednich podrozdziałach, który został użyty do przeprowadzenia symulacji przewidywania o trzydzieści minut na podstawie danych testowych z trzech miesięcy.

```
=== Summary ===

Correctly Classified Instances         351               53.8344 %
Incorrectly Classified Instances       301               46.1656 %
Kappa statistic                          0.0811
Mean absolute error                      0.3198
Root mean squared error                  0.4204
Relative absolute error                 94.7006 %
Root relative squared error            102.2551 %
Total Number of Instances              652

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,508     0,422     0,523       0,508   0,515      0,086    0,576     0,560     up
               0,574     0,497     0,551       0,574   0,563      0,078    0,570     0,583     down
               0,000     0,000     0,000       0,000   0,000      0,000    0,491     0,008     no
Weighted Avg.  0,538     0,457     0,534       0,538   0,536      0,081    0,572     0,568

=== Confusion Matrix ===

   a   b   c   <-- classified as
 158 153   0 |   a = up
 143 193   0 |   b = down
   1   4   0 |   c = no
```

Rysunek 5.8 Podsumowanie wytrenowanego modelu do predykcji o 30 min
Źródło: opracowanie własne

Otrzymany model ma podobne wyniki do modelu przygotowanego do predykcji o 5 min. Precyzja tego modelu wynosi 52.3% dla ruchów w górę i 55.1% dla ruchów w górę. Czułość modelu wynosi 50.8% dla ruchów w górę i 57.4% dla ruchów w dół. Wartość poprawnie sklasyfikowanych instancji wynosi 53.8%.

```
=== Summary ===

Correctly Classified Instances        3167               50.7776 %
Incorrectly Classified Instances      3070               49.2224 %
Kappa statistic                          0.0262
Mean absolute error                      0.3363
Root mean squared error                  0.4333
Total Number of Instances             6237

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,476    0,450    0,517      0,476    0,496      0,026    0,514     0,513     up
               0,551    0,523    0,500      0,551    0,524      0,027    0,515     0,493     down
               0,000    0,000    0,000      0,000    0,000      0,000    0,457     0,009     no
Weighted Avg.  0,508    0,482    0,504      0,508    0,505      0,026    0,514     0,498

=== Confusion Matrix ===

    a    b    c   <-- classified as
 1493 1644    0 |    a = up
 1365 1674    0 |    b = down
   31   30    0 |    c = no
```

Rysunek 5.9 Podsumowanie ogólne symulacji 3-miesięcznej próby modelu predykcji o 30 min
Źródło: opracowanie własne

Na powyższym rysunku widać, że przy symulacji użycia modelu na danych z trzech miesięcy dla predykcji o trzydzieści minut zanotowano spodziewany spadek skuteczności względem wyników treningowych. Jednak widoczny jest również niewielki wzrost skuteczności modelu względem wyników tego badania na danych predykcji o pięć minut. Precyzja określenia ruchów w górę wynosi 51.7%, a ruchów w dół 50%. Czułość ruchów w górę wyniosła 47.6%, a ruchów w dół 55.1%.

miesiąc	Precision UP	Recall UP	Precision DOWN	Recall DOWN	Precision NO	Recall NO
marzec	0,498	0,498	0,505	0,515	0	0
kwiecień	0,504	0,448	0,509	0,577	0	0
maj	0,543	0,481	0,49	0,559	0	0

Tabela 5.5 Rezultaty symulacji predykcji o 30 min z podziałem na miesiące
Źródło: opracowanie własne

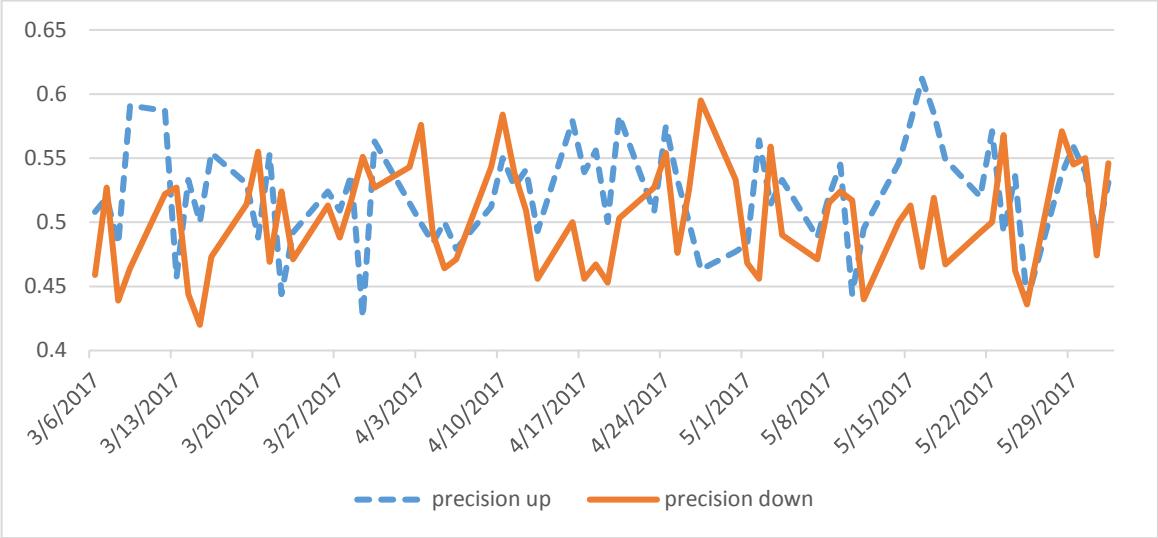
Tabela 5.10 przedstawia wyniki badania w podziale na miesiące, gdzie zróżnicowanie wyników zarówno w odniesieniu do precyzji jak i czułości niewielkie, ale zauważalne.

tydzień	Precision UP	Recall UP	Precision DOWN	Recall DOWN	Precision NO	Recall NO
1-03	0,525	0,519	0,443	0,509	0	0
2-03	0,477	0,433	0,477	0,534	0	0
3-03	0,496	0,485	0,504	0,525	0	0
4-03	0,496	0,515	0,533	0,524	0	0
1-04	0,456	0,464	0,5	0,612	0	0
2-04	0,51	0,442	0,529	0,602	0	0
3-04	0,541	0,441	0,461	0,576	0	0
4-04	0,508	0,546	0,551	0,522	0	0

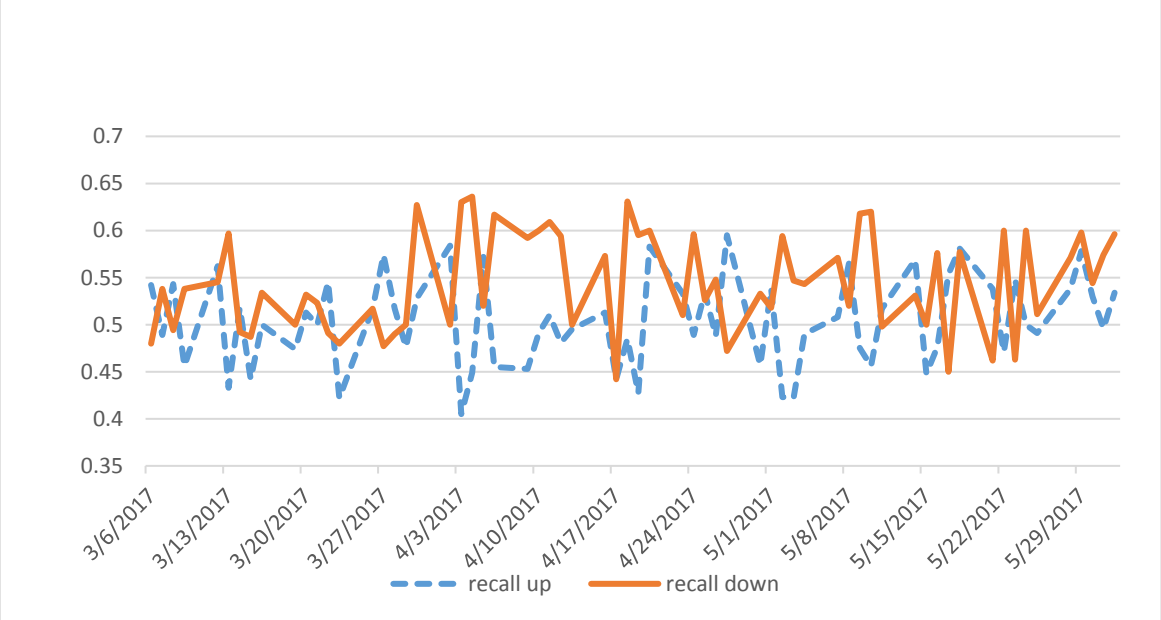
1-05	0,527	0,478	0,488	0,546	0	0
2-05	0,498	0,409	0,473	0,565	0	0
3-05	0,584	0,504	0,455	0,52	0	0
4-05	0,535	0,483	0,513	0,572	0	0
5-05	0,558	0,525	0,538	0,584	0	0

Tabela 5.6 Rezultaty symulacji predykcji o 5 min z podziałem na tygodnie
Źródło: opracowanie własne

Z tabeli prezentującej wyniki w podziale na tygodnie widać wyraźniej poprawę wyników precyzji w porównaniu do predykcji o pięć minut. Zakres precyzji dla ruchów w górę wynosi od 46% do 58% oraz dla ruchów w dół od 44% do 54%. Natomiast zakres czułości dla ruchów w górę wynosi od 41% do 55%, a dla ruchów w dół od 51% do 61%.



Rysunek 5.10 Wykres wyników precyzji predykcji o 30 min z podziałem na dni
Źródło: opracowanie własne



Rysunek 5.11 Wykres wyników czułości predykcji o 30 min z podziałem na dni
Źródło: opracowanie własne

Na rysunkach z podziałem na dni widać największe zróżnicowanie wyników. Precyzja modelu mieści się w zakresie od 45% do 59% dla ruchów w górę oraz dla ruchów w dół od 45% do 57%. Czułości modelu przy podziale na dni mieszczą się w zakresie od 45% do 63% dla ruchów w górę i podobnie dla ruchów w dół. Widoczne jest nieznaczne przesunięcie do góry zakresów precyzji i skuteczności względem wyników przy predykcji o pięć minut.

5.4.3 Test - prognoza o godzinę

Poniżej na rysunku przedstawiony jest wynik uzyskanego modelu bazującego na parametrach opisanych w poprzednich podrozdziałach, który został użyty do przeprowadzenia symulacji przewidywania o jedną godzinę na podstawie danych testowych z trzech miesięcy.

```
=== Summary ===

Correctly Classified Instances         539               50.8491 %
Incorrectly Classified Instances       521               49.1509 %
Kappa statistic                          0.0244
Mean absolute error                      0.3355
Root mean squared error                  0.4302
Relative absolute error                 98.8848 %
Root relative squared error            104.5961 %
Total Number of Instances             1060

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,516     0,490     0,503       0,516   0,510      0,027    0,518     0,495     down
               0,508     0,486     0,514       0,508   0,511      0,023    0,519     0,535     up
               0,000     0,000     0,000       0,000   0,000      0,000    0,533     0,009     no
Weighted Avg.  0,508     0,484     0,505       0,508   0,507      0,024    0,519     0,512

=== Confusion Matrix ===

   a   b   c   <-- classified as
 268 251   0 |   a = down
 262 271   0 |   b = up
   3   5   0 |   c = no
```

Rysunek 5.12 Podsumowanie wytrenowanego modelu do predykcji o 1 h
Źródło: opracowanie własne

W przypadku tego modelu można zauważyć już spadek skuteczności względem dwóch poprzednich modeli użytych do symulacji. Precyzja predykcji ruchów w górę wyniosła 51.4% przy czułości równej 50.8%, a precyzja predykcji ruchów w dół wyniosła 50.3% przy czułości 51.6%. Sklasyfikowanych prawidłowo zostało 49.1% obiektów.

```
=== Summary ===

Correctly Classified Instances        1571               50.3849 %
Incorrectly Classified Instances      1547               49.6151 %
Kappa statistic                          0.0187
Mean absolute error                      0.3396
Root mean squared error                  0.4356
Total Number of Instances             3118

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,570    0,549    0,497      0,570    0,531      0,021    0,504     0,489     down
                 0,449    0,432    0,513      0,449    0,479      0,017    0,498     0,499     up
                 0,000    0,000    0,000      0,000    0,000      0,000    0,448     0,020     no
Weighted Avg.    0,504    0,485    0,500      0,504    0,500      0,019    0,500     0,490

=== Confusion Matrix ===

   a   b   c   <-- classified as
 867 653   0 |   a = down
 865 704   0 |   b = up
  13  16   0 |   c = no
```

Rysunek 5.13 Podsumowanie ogólne symulacji 3-miesięcznej próby modelu predykcji o 1 h
Źródło: opracowanie własne

W tym przypadku również zauważalny jest spadek precyzji i czułości względem tej w podsumowaniu modelu. Dla przewidywania ruchów w górę precyzja wyniosła 51.3%, a czułość 44.9%. Dla przewidywania ruchów w dół precyzja wyniosła 49.7%, a czułość 57%.

miesiąc	Precision UP	Recall UP	Precision DOWN	Recall DOWN	Precision NO	Recall NO
marzec	0,506	0,462	0,522	0,577	0	0
kwiecień	0,498	0,441	0,493	0,531	0	0
maj	0,533	0,432	0,481	0,599	0	0

Tabela 5.7 Rezultaty symulacji predykcji o 1 h z podziałem na miesiące
Źródło: opracowanie własne

W powyższym zauważalne jest niewielkie zróżnicowanie wyników ze względu na konkretne miesiące. Wyniki oscylują w okolicach 50% tak jak w poprzednich badaniach.

tydzień	Precision UP	Recall UP	Precision DOWN	Recall DOWN	Precision NO	Recall NO
1-03	0,524	0,473	0,513	0,572	0	0
2-03	0,491	0,442	0,492	0,546	0	0
3-03	0,485	0,42	0,504	0,582	0	0
4-03	0,522	0,528	0,596	0,508	0	0
1-04	0,494	0,425	0,445	0,542	0	0
2-04	0,459	0,496	0,486	0,423	0	0
3-04	0,502	0,512	0,492	0,598	0	0
4-04	0,541	0,454	0,543	0,579	0	0
1-05	0,515	0,42	0,503	0,605	0	0
2-05	0,506	0,436	0,452	0,625	0	0
3-05	0,595	0,419	0,434	0,613	0	0
4-05	0,529	0,502	0,538	0,588	0	0
5-05	0,521	0,475	0,513	0,517	0	0

Tabela 5.8 Rezultaty symulacji predykcji o 1 h z podziałem na tygodnie
Źródło: opracowanie własne

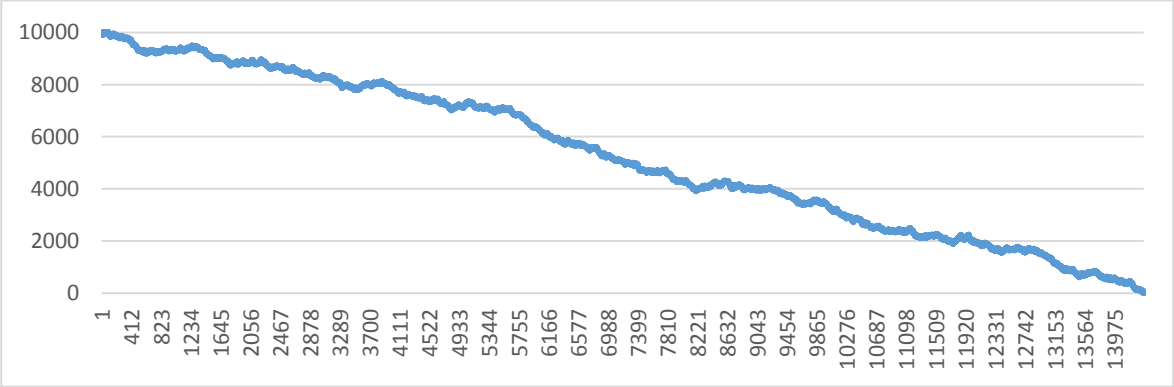
Na podziale tygodniowym można zauważyć, że większość wyników precyzji ruchów w górę mieści się w zakresie od 49% do 53%. Natomiast dla ruchów w dół większość wyników dotyczących precyzji należy do zakresu to od 45% do 54%. Czułość decyzji dla ruchów w górę należy do zakresu do 42% do 51%, a dla ruchów w dół od 51% do 61%. Testy z podziałem na dni nie zostały przeprowadzone dla predykcji o godzinę, ponieważ zbiory w podziale na pojedyncze dni były zbyt małe by wyniki otrzymane na tak podzielonych danych były miarodajne.

5.5 Symulacja gry na opcjach binarnych.

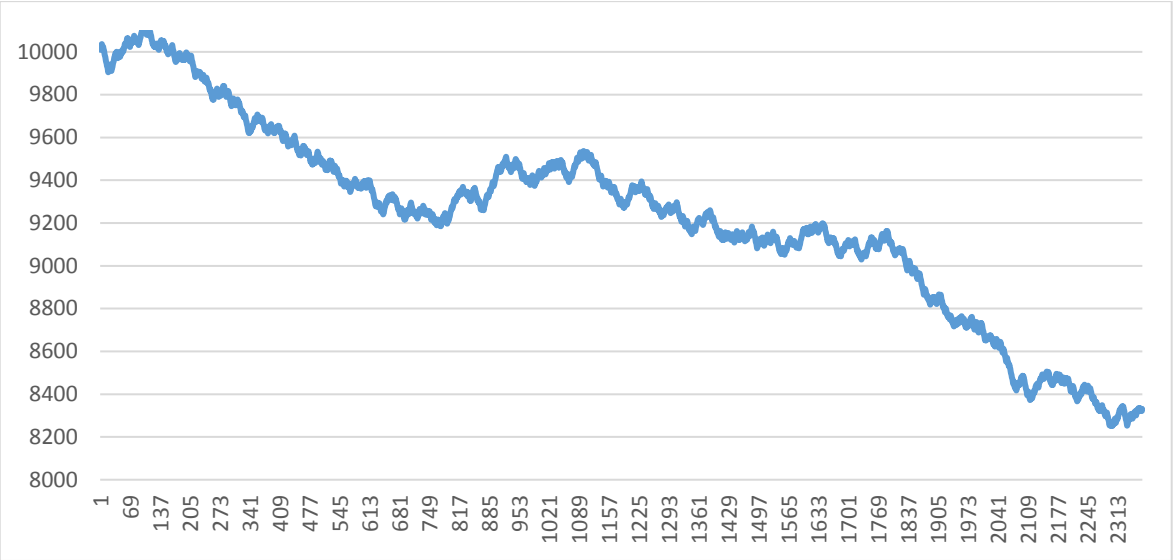
Do przeprowadzenia symulacji wykorzystany został miesięczny zbiór danych historycznych z maja 2017 roku. Symulacja została przeprowadzona w sposób imitujący grę na opcjach binarnych. Oznacza to że na potrzeby symulacji przyjęto:

- zwrot z poprawnie prognozowanej opcji wynoszący 90% kwoty zainwestowanej,
- każda prognoza wykonana przez algorytm oznaczała podjęcie inwestycji o wartości 10 PLN,
- okres inwestycyjny wynoszący jeden miesiąc.

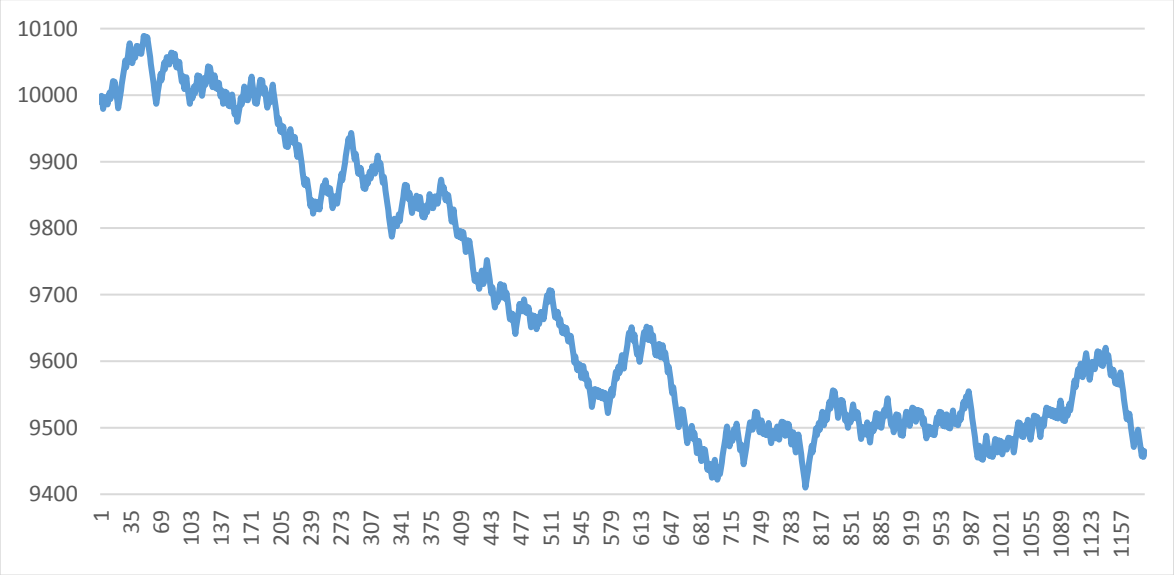
Zgodnie ze sposobem działania opcji binarnych i powyższymi założeniami. Każda prawidłowa decyzja oznacza zwrot zainwestowanej kwoty powiększonej o 90%, co przy inwestycji wynoszącej 10 PLN oznacza dodanie 9PLN na konto inwestora. W przypadku decyzji błędnej 10PLN jest odbierane z konta. W przypadku, gdy cena z czasem wygaśnięcia okresu prognozy jest równa cenie w chwili podjęcia inwestycji z konta nie jest pobierana żadna kwota. Symulacja prowadzona była do momentu wyczerpania się kapitału na koncie lub zakończenia się okresu inwestycyjnego. Oś Y wykresów przedstawia kapitał, natomiast oś X ilość decyzji.



Rysunek 5.14 Wykres stanu kapitału w stosunku do liczby podjętych decyzji – prognoza o 5 minut
Źródło: opracowanie własne



Rysunek 5.15 Wykres stanu kapitału w stosunku do liczby podjętych decyzji – prognoza o 30 minut
Źródło: opracowanie własne



Rysunek 5.16 Wykres stanu kapitału w stosunku do liczby podjętych decyzji – prognoza o 1 godzinę
Źródło: opracowanie własne

Biorąc pod uwagę wynik wcześniej wykonanych testów spadek kapitału jest spodziewanym rezultatem symulacji. Zasady opcji binarnych są skonstruowane w taki sposób, że wygrana nie jest w stanie pokryć straty dlatego algorytm ze skutecznością w okolicach 50% przynosi straty. Jedyną symulacją, która doprowadziła do straty całego kapitału jest symulacja o pięć minut. Jest to wynikiem dużej ilości podjętych decyzji oraz najniższej skuteczności modelu ze wszystkich trzech symulacji. Kolejne dwa wykresy pokazują, że podobnie jak w testach, symulacje prognozy o trzydzieści minut i o godzinę dają lepszy wynik. Na wykresach wyraźnie widoczne są okresy, kiedy algorytm prognozuje

prawidłowo kierunek ruchu ceny, wtedy wykres porusza się do góry. Oraz okresy błędnych decyzji, kiedy wykres porusza się do dołu. Ruchy w dół są gwałtowniejsze jednak jest to spowodowane wcześniej wspomnianą specyfiką zasad opcji binarnych.

5.6 Podsumowanie testów i symulacji

Wynikiem badań jest wybrany optymalny algorytm uczenia maszynowego dający najlepsze rezultaty z badanych algorytmów. Dla wszystkich testów wyniki wyszły bardzo zbliżone do siebie. Precyzja i czułość prognozy ruchów w każdym z przypadków oscylowała w okolicach 50%. Zróżnicowanie wyników przy podziale na miesiące i tygodnie nie było duże, ponieważ sięgało nie więcej niż 4% w przypadku precyzji i 6% w przypadku czułości prognozy kierunku ruchu. Większe zróżnicowanie było widoczne przy podziale na poszczególne dni. Nie jest to zaskakującym rezultatem ze względu na charakter rynku giełdowego którym rządzą trendy, a częste wydarzenia polityczne powodują niespodziewane ruchy. Dodatkowo rozbicie zbioru danych na mniejsze części sprzyjało większym wahaniom wyników, ponieważ im mniejszy zbiór tym większy procentowy wpływ ma pojedyncza decyzja na wyniki. Tylko model prognozy o pięć minut był w stanie przewidywać skutecznie braki ruchów rynku, czyli pozycje „Precision NO” i „Recall NO”. Jest to logicznym zachowaniem modelu uczenia maszynowego, ponieważ dostarczone dane do predykcji o pięć minut były najbardziej szczegółowe i pozwalały na przewidzenie takich wyjątkowych zachowań na rynku jak brak ruchu ceny. Jednak tak szczegółowe dane nie są tak skuteczne przy przewidywaniu o trzydzieści minut czy o godzinę, dlatego należało przebudować model do takich predykcji.

Wyniki symulacji zgadzały się z wynikami testów, wstępnie założony kapitał zmniejszał się wraz z ilością podejmowanych decyzji. Na wykresach wyraźnie widoczne były momenty, gdy model przewidywał kierunki ruchów prawidłowo co skutkowało wzrostem kapitału. Oraz spadki kapitału oznaczające błędne decyzje modelu. Wynik w okolicach 50% nie jest wystarczającą by podejmować sukcesywne inwestycje na opcjach binarnych.

6. Podsumowanie i wnioski

Celem pracy było wybranie optymalnego algorytmu uczenia maszynowego oraz przeprowadzenie symulacji gry na opcjach binarnych z użyciem wybranego modelu, czyli przeprowadzanie prognozy kierunku kursów giełdowych w określonych przedziałach czasowych.

Przedstawione zostały podstawowe informacje dotyczące rynków giełdowych, opcji binarnych, analizy technicznej i wskaźników. Opisane zostało uczenie maszynowe i wybrane algorytmy. Zaprezentowane zostały narzędzia użyte do wygenerowania danych i przeprowadzenia badań. Krok po kroku opisane zostały etapy pracy. W przedostatnim rozdziale zaprezentowane zostały wyniki badań oraz symulacji.

Biorąc pod uwagę jeden z elementów zakresu pracy, czyli przeprowadzenie symulacji gry na opcjach binarnych, skupiono się na prognozie kierunku kursu w określonym przedziale czasowym. W związku z tym wybrany został algorytm drzew losowych jako optymalny, dający najlepsze efekty dla użytego zestawu danych w prognozie kierunku kursu. Do badań użyte zostały rzeczywiste historyczne dane giełdowe, z których z pomocą programu napisanego w języku Java zostały wygenerowane wartości wskaźników analizy technicznej oraz sygnały stanów rynkowych. Wyniki badań zostały zaprezentowane w formie tabel i wykresów.

Można stwierdzić, że model nie zdołał przewidzieć kierunku kursu, na poziomie który nadawałby się do inwestowania na opcjach binarnych. Użycie takiego algorytmu przy inwestowaniu w opcje binarne nie byłoby opłacalne. Zakres czułości wąchający się w okolicach 50% jest wynikiem na średnim poziomie, jednak połowa instancji była klasyfikowana prawidłowo przy tak losowym i zawierającym wiele szumów źródle danych jak giełda forex. Również zwiększenie przedziału czasowego w którym były wykonywane symulacje jak się spodziewano nieznacznie wpłynęła na polepszenie wyników. Było to spowodowane mniejszą ilością szumów na wyższych interwałach czasowych. Świadczy to o tym, że model działa jednak nie ma silnego powiązania pomiędzy danymi dostarczonymi w zbiorach a kierunkiem kursu, lub dostarczone dane okazały się zbyt zaszumione.

Obszar dotyczący wykrywania zależności w tego typu dużych zbiorach zaszumionych danych jest nowy jednak ma duży potencjał. Dodatkowo warto do zbioru danych wprowadzić informacje o budujących się formacjach świecowych na rynku w czasie rzeczywistym. Takie wykrywanie wzorców świecowych również jest dobrym materiałem

do wykorzystania uczenia maszynowego. Jest to bardzo obszerny temat, nadający się na kolejną pracę magisterską. Wprowadzenie ograniczenia podejmowanych decyzji do sytuacji, gdy cena znajduje się w ekstremalnych wartościach wskaźników prawdopodobnie spowodowałoby spadek szumów. Budowany model bazował jedynie na analizie technicznej, natomiast dodanie elementów analizy fundamentalnej mogłoby nauczyć model reakcji rynku na wydarzenia polityczne.

7. Literatura

1. D. R. Lambert, *Commodity Channel Index: Tool for Trading Cyclic Trends, Commodities Magazine* 1980, Stocks & Commodities V, s 120-122
2. G. Appel, E. Dobson, *Understanding MACD*, Greenville, Traders Press, 2008
3. J. Bollinger, *Bollinger on Bollinger Bands*, McGraw-Hill New York 2001
4. J.J Murphy, *Analiza techniczna rynków finansowych*, WIG-PRESS Warszawa 1999
5. M.C Thomsett, *Mastering Fundametnal Analysys*, Dearborn Financial Publishing 1998
6. P. Cichosz, *Systemy uczące się, Wydawnictwa Naukowo-Techniczne*, Warszawa, 2000
7. R. D. Edwards, J. Magee, W.H.C. Bassetti, *Technical analysis of stock trends*, CRC Press New York 2007
8. S. Nison, *Japanise candlestick charing techniques*, Prentice Hall Press USA 2001
9. S. Shalev-Shwartz, S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press 2014
10. Bank for international settlement, *Foreign exchange turnover in April 2013: preliminary global results*, Witryna internetowa http://www.bis.org/publ/rpfx13fx.pdf stan na na 26.07.2017
11. C. Lai, M. J.T. Reinders, L. Wessels, *Random Subspace Method for multivariate feature selection*, Witryna internetowa http://isplab.tudelft.nl/sites/default/files/Lai05a.pdf dostęp na stan na 26.07.2017
12. Dokumentacja techniczna TA4J. Witryna internetowa https://github.com/mdeverdelhan/ta4j/wiki stan na 25.07.2017
13. G. Louppe, *Understanding random forests from theory to practice*, University of Liège, Witryna internatowa https://arxiv.org/pdf/1407.7502.pdf stan na 12.07.2017
14. J. Kuepper, *Basics Of Technical Analysis,* Witryna internetowa http://www.investopedia.com/university/technical stan na 10.07.2017
15. J. Mazurek, *Czym jest rynek Forex?*, Witryna internetowa http://www.bankier.pl/wiadomosc/Czym-jest-rynek-Forex-1797860.html stan na 27.06.2017

16. Majors, Minors & Exotic Currency Pairs, Witryna internetowa http://www.sharptrader.com/new-to-trading/forex/majors-minors-exotic-currency-pairs/ stan na 27.06.2017
17. WEKA Manual for Version 3-9-1. Witryna internatowa http://www.cs.waikato.ac.nz/ml/weka/documentation.html stan na 26.07.2017
18. Zespół developerski Ceneo.pl, *Machine Learning w Ceneo*, Witryna internetowa https://devstyle.pl/2017/08/09/machine-learning-w-ceneo/ stan na 2.09.2017

8. Spis ilustracji

9. Spis tabel