**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**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| POLITECHNIKA BIAŁOSTOCKA | | |  | | | Nr albumu studenta  90563. | |
| Wydział Informatyki | | | Studia stacjonarne  II stopnia inżynierskie | | | Rok akademicki  2015/2016 | |
| Kierunek studiów  Informatyka | |
| Katedra Oprogramowania | | |  | | | 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 | | | | | | | |
| .........................................................................  *Imię i nazwisko, stopień/ tytuł promotora - podpis* | | | | ......................................................................................  *Imię i nazwisko kierownika katedry - podpis* | | | |
| .....................................................*Data wydania tematu pracy dyplomowej*  *- podpis promotora* | | ...................................................  *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* | |

**Karta dyplomowa**

Thesis topic:

Web application for supporting sport club activity.

SUMMARY

The aim of the project was creating a web application for supporting sport club activity. The motivation behind choosing the thesis was my membership in Parkour Białystok sport club to which the application is dedicated. This project allows to manage various articles, members, users, photo gallery and competitions organized by sport club.

This application was created using modern technologies, these are most important: Spring Framework and based-on child projects like Spring Boot, Thymeleaf, JQuery, Bootstrap.

This thesis consists of six chapters. The first chapter is an introduction. The second presents similar solutions. The next one presents and describes technologies which were used to create application. All diagrams and schemas created during planning process are described in chapter four. Chapter five is most important as it contains the technical and implementation details. These include a description of URL mapping, graphic interface, forms, server implementation, configuration and security aspect. The last chapter is a summary of work.

Załącznik nr 4 do „Zasad postępowania przy przygotowaniu i obronie pracy dyplomowej w PB”

Załącznik nr 4 do „Zasad postępowania przy przygotowaniu i obronie pracy dyplomowej w PB”

Maciej Ziniewicz Białystok, dnia………………………

imię i nazwisko studenta

90563

nr albumu

Informatyka stacjonarne

kierunek i forma studiów

dr inż. Jerzy Krawczuk

promotor pracy dyplomowej

**OŚWIADCZENIE**

Przedkładając w roku akademickim 2015./2016. Promotorowi dr inż. Jackowi Grekowowi pracę dyplomową pt.: Aplikacja webowa wspomagająca działanie klubu sportowego dalej zwaną pracą dyplomową,

**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.).

Białystok, dnia………………………

imię i nazwisko studenta

................................................

nr albumu

...............................................

kierunek i forma studiów

...............................................

promotor pracy dyplomowej

OŚWIADCZENIE

Przedkładając w roku akademickim 2015./2016. Promotorowi ………………………….……………………….. pracę dyplomową pt.: ...............…………………………………………………….……………………………………….. ………………….………………………..………………………………………..……………………………...………….…….…………………………………….........................…………………………………, dalej zwaną pracą dyplomową,

**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/am ich autora, tytuł, miejsce i rok publikacji oraz stronę, z której pochodzą powoływane fragmenty, ponadto w pracy dyplomowej zamieściłem/am 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,

# Spis treści

[Spis treści](#_Toc490926712)

[1. Wstęp](#_Toc490926713)

[2. Predykcja rynków](#_Toc490926714)

[2.1 Rynek Forex](#_Toc490926715)

[2.1.1 Pary walutowe](#_Toc490926716)

[2.2 Analiza fundamentalna](#_Toc490926717)

[2.3 Analiza techniczna](#_Toc490926718)

[2.3.1 Świece japońskie](#_Toc490926722)

[2.3.2 Analiza trendu](#_Toc490926723)

[2.3.3 Wskaźniki](#_Toc490926724)

[2.4 Uczenie maszynowe](#_Toc490926725)

[2.4.1 Pojęcie uczenia się w kontekscie maszynowym](#_Toc490926726)

[2.4.2 Podstawowe pojęcia](#_Toc490926727)

[2.4.3 Przedstawienie badanych algorytmów](#_Toc490926728)

[3. Opis wykorzystanych narzędzi](#_Toc490926729)

[3.1 Technical Analysis for Java (Ta4j)](#_Toc490926730)

[3.1.1 Podstawy korzystania z TA4j](#_Toc490926731)

[3.2 Weka](#_Toc490926732)

[3.2.1 Wstępne przetworzenie danych](#_Toc490926733)

[3.2.2 Klasyfikacja danych](#_Toc490926734)

[3.2.3 Pozostałe funkcjonalności](#_Toc490926735)

[4. Metodologia](#_Toc490926736)

[4.1 Charakter badań](#_Toc490926737)

[4.2 Ogólny przebieg badań](#_Toc490926738)

[4.3 Omówienie użytych danych](#_Toc490926739)

[4.4 Omówienie procesu obróbki danych do sygnałow](#_Toc490926740)

[4.4.1 Tworzenie świec japońskich](#_Toc490926741)

[4.4.2 Generowanie sygnałów](#_Toc490926742)

[4.4.3 Wygenerowane dane](#_Toc490926743)

[4.5 Omówienie procesu badań](#_Toc490926744)

[4.5.1 Wznaczenie i filtrowanie atrybutów](#_Toc490926745)

[4.5.2 Klasyfikacja](#_Toc490926746)

[4.5.3 Testowanie](#_Toc490926747)

[5. Rezultat badań](#_Toc490926748)

[5.1 Klasyfikator](#_Toc490926749)

[5.2 Atrybuty](#_Toc490926750)

[5.3 Ustawienia wskaźników](#_Toc490926751)

[5.4 Rezultaty predyckja 5 min](#_Toc490926752)

[5.5 Rezultaty predykcja 30min](#_Toc490926753)

[5.6 Rezultaty predykcja 1h](#_Toc490926754)

[6. Wnioski](#_Toc490926755)

[7. Bibliografia](#_Toc490926756)

# 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 analizować te informacje i wspomagać proces decyzyjny bądź nawet podejmować decyzje inwestycyjne za człowieka. Niestety rynek charakteryzuje się dużą ilością hałasów i niestabilnością przez co prognoza jest bardzo trudnym zadaniem.

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 instrumentu 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.

Praca podzielona jest na 6 rozdziałów. W rozdziale drugim zostanie przeprowadzony 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 spekulacji. Ostatni szósty rozdział stanowi podsumowanie, opisuje wnioski płynące z pracy magisterskiej i sugeruje dalsze kierunki badań.

# 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.

## 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[[1]](#footnote-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 warty 4.23 złotego, czy wartość euro do dolara 1.2 oznacza, że jeden euro jest warty 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[[2]](#footnote-2):

1. **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 http://www.bis.org/publ/rpfx13fx.pdf (stan na 27.06.2017)

1. **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

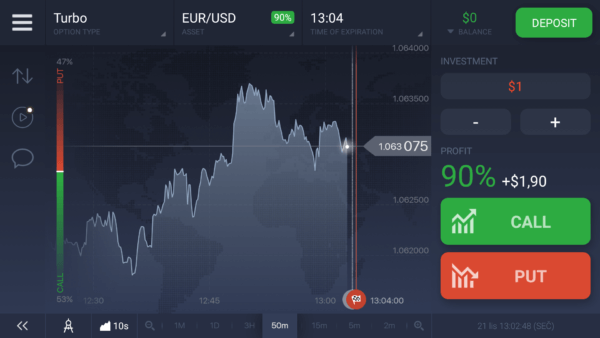
1. **Egzotyczne** - są to pary walutowe składające się z jednej z głównych bardziej znaczących walut do walut które są mniej znaczące ekonomicznie w skali globalnej. Handel z użyciem tych par walutowych stanowi najmniejszą część rynku Forex.

|  |  |
| --- | --- |
| 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

## 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 wzrośnie czy niższa z momentem wygaśnięcia opcji, czyli z końcem ustalonego przedziału czasowego. 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 wartość ceny nie była zgodna 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, niezabezpieczeniu otwartych opcji może skutkować stratą całości kapitału.



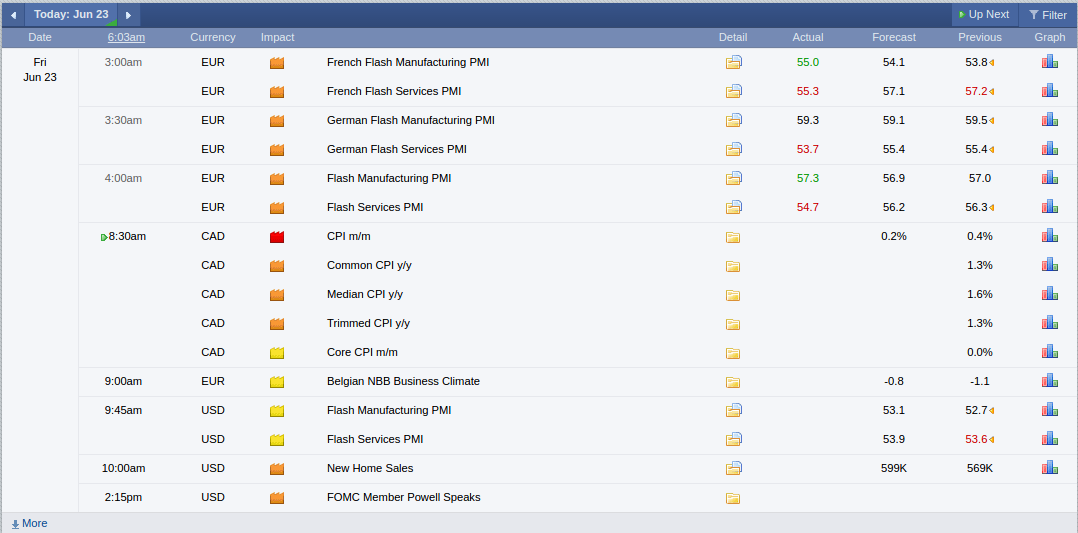
Rysunek 2.1 Widok aplikacji IQ Options brokera opcji binarnych  
Źródło: https://xbinop.com/pl/opinia/iqoption/ (stan na 16.09.2017)

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 w górę natomiast wybór „PUT” jest to ruch 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.

## 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[[3]](#footnote-3). Tego rodzaju analizy są stosowane głownie w długoterminowych inwestycjach oraz częściej to pojęcie jest spotykane przy rynkach papierów wartościowych, 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.

## 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]](#footnote-4). Przede wszystkim w analiza techniczna polega na odczytywaniu informacji z wykresów instrumentów giełdowych, w tym przypadku będzie mowa o walutach. Na podstawie obserwacji zachowań ceny m.in: nagłych skoków, utworzonych formacji możliwe jest by określić przyszły trend. Cała koncepcja analizy technicznej opiera się o trzy przesłanki:

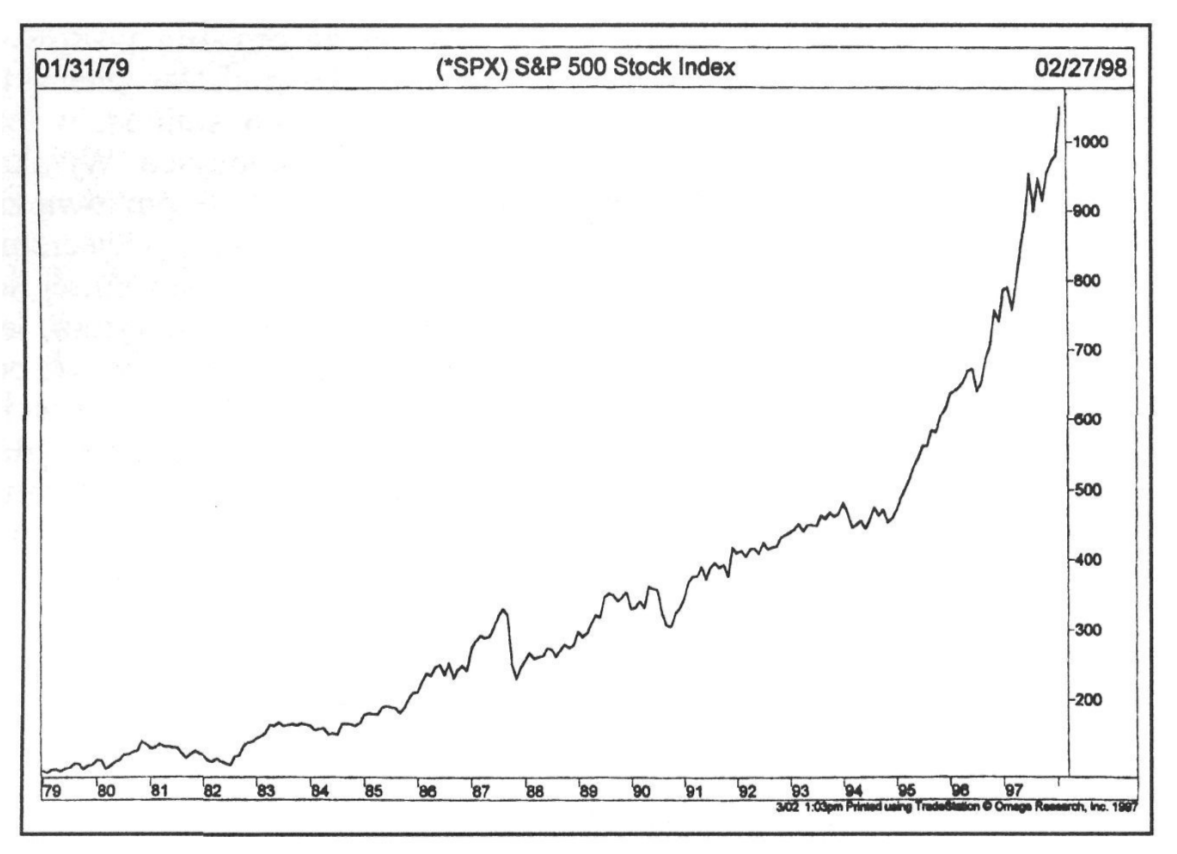
* Rynek dyskontuje wszystko
* Ceny podlegają trendom
* Historia się powtarza[[5]](#footnote-5)

### **Rynek dyskontuje wszystko**

Ta przesłanka jest podstawą analizy technicznej z której wynikają kolejne[[6]](#footnote-6). 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.

### **Ceny podlegają trendom**

Kolejne podstawowe założenie technicznej mówi, że ceny podlegają trendom krótko, średnio, długo terminowym. Rynek wykazuje tendencje do kontynuowania trendu niż do jego zmiany[[7]](#footnote-7). Innymi słowy duża część analizy technicznej sprowadza się do tego by badać trendy oraz wykrywać ich zmiany 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

### **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.

### 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 nazywającego się 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]](#footnote-8).

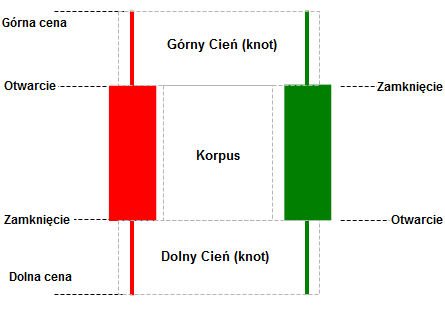
Wykres świecowy daje największą ilość informacji, w porównaniu do dwóch innych również często używanych, czyli liniowego oraz słupkowego.



Rysunek 2.4 Rodzaje wykresów  
Źródło: opracowanie własne

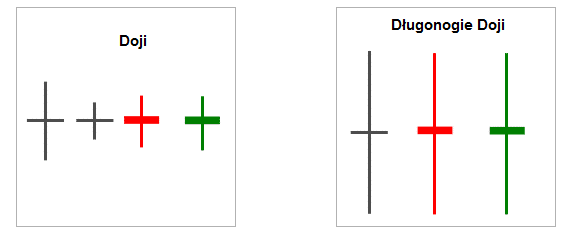
Tak jak widać na powyższym rysunku (rys. 2.3) wykres liniowy dostarcza najmniejszą ilość informacji dla inwestora. Łączy on ceny zamknięcia z każdego przedziału czasowego to znaczy, że inwestor otrzymuje informacje o cenie otwarcia, czyli cenie początkowej w określonym przedziale czasowym oraz cenie 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 które dostarczane są przez świece są jej cienie, a informują one o ekstremach ceny występujących w czasie przedziału czasowego prezentowanego przez świecę. Cień górny informuje o maksymalnej cenie jak wystąpiła podczas przedziału, a cień dolny informuje o cenie minimalnej[[9]](#footnote-9).



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

Wygląd pojedynczych świec może również stanowić dodatkową informację, ale również sposób w jaki ustawiają się świece kolejno po sobie jest wartościową informacją dla inwestora, nazwane jest to formacjami. 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[[10]](#footnote-10). Doji mogą różnić się długością cienia.



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

### 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”[[11]](#footnote-11), 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 ze względu sytuację, kiedy cena nie podąża ani do góry, ani do dołu.



Rysunek 2.8 Trend boczny  
Źródło: opracowanie własne

W trendach wyróżnia się nie tylko 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 wystąpić krótszy trend średnio okresowy boczny bądź spadkowy który obniży cenę następnie jak się skończy cena będzie nadal się poruszała zgodnie z trendem głównym[[12]](#footnote-12).

### Wskaźniki

Wskaźniki używane są podstawowym narzędziem 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 przy podejmowaniu decyzji, mogą przynieść odwrotny efekt dlatego nie należy traktować ich jako podstawę do decyzji oraz korzystanie ze zbyt dużej ilości wskaźników może wprowadzić w błąd.

Wskaźniki możemy podzielić na dwa typy[[13]](#footnote-13):

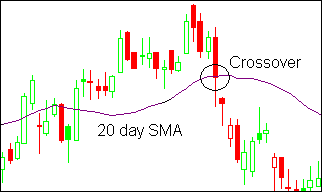
1. Wiodące (ang. leading) – wskaźniki poprzedzające ruchy ceny oraz próbujące je przewidzieć. Przydają się podczas trendów bocznych, gdy rynek nie ma widocznych trendów.
2. Opóźnienia (ang. lagging) – wskaźniki podążające za ruchem ceny, śledzące ją. Służą najczęściej jako potwierdzenia wspomagają decyzje.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

#### Oscylatory

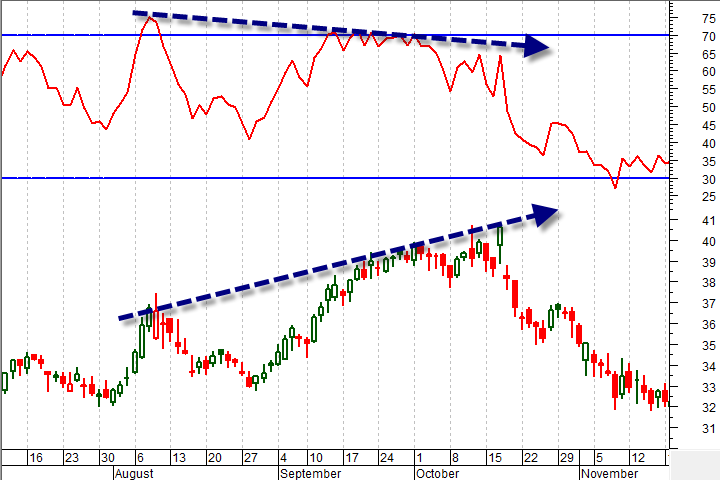
Szczególnej uwagi wymagają oscylatory. Wskaźniki należące do tej rodziny są pomocnicze przy analizie trendu, którego skuteczność jest największa kiedy nie ma dużych ruchów cen. Konstrukcja wszystkich oscylatorów jest podobna, jest to rysowany na bieżąco wykres o osi na poziomie zero która dzieli wykres na górną i dolną połowę lub o maksymalnej wartości 100 i minimalnej 0. Oscylatory sygnalizują możliwość zmiany kierunku ruchu ceny w momencie zbliża się wykresu do ekstremalnych wartości oscylatora. Gdy wartość oscylatora jest blisko maksimum mowa jest o wykupieniu rynku, natomiast gdy jego wartość zbliża się do minimum mówimy o wysprzedaniu rynku[[14]](#footnote-14).

Najczęściej poszukiwane na oscylatorach są przecięcia linii (ang. crossovers) oraz dywergencje (ang. divergence). Przecięcia linii używane są często w odniesieniu do średnich ruchomych czy wstęg Bollingera. Wskaźniki te można zaliczyć do wskaźników opóźnienia, ponieważ śledzą one ruch ceny. 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.[[15]](#footnote-15) Poniżej przykład przecięcia linii dwudziestodniowej średniej kroczącej.



Rysunek 2.9 Przecięcie linii średniej kroczącej  
Źródło: http://www.investopedia.com/terms/c/crossover.asp (dostęp 11.07.2017)

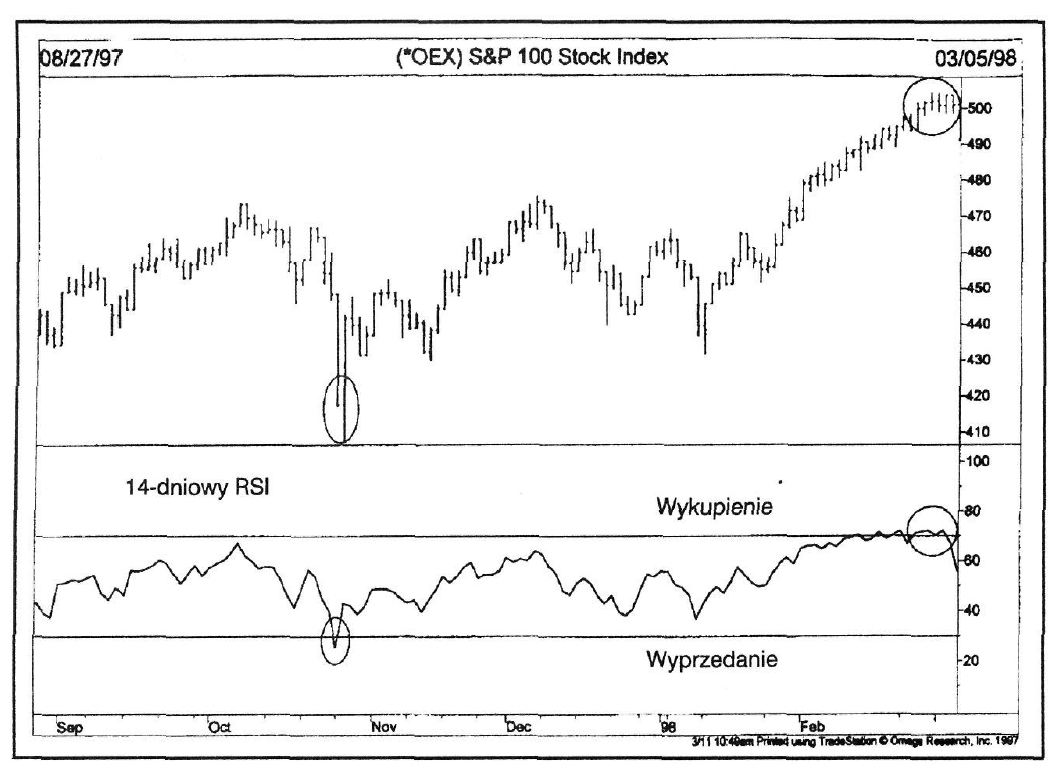
Dywergencje są to sytuacje, kiedy wskaźnik i wykres ceny podążają w przeciwnych kierunkach. Przykładowo, gdy cena rośnie, a wartość wskaźnika spada może być to interpretowane jako sygnał odwrócenia trendu. Dywergencja jest silniejsza im większa jest rozbieżność między wskaźnikiem a wykresem[[16]](#footnote-16).



Rysunek 2.10 Dywergencja  
Źródło: https://www.learningmarkets.com/how-to-trade-bullish-and-bearish-technical-divergences/ (dostęp 11.07.2017)

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

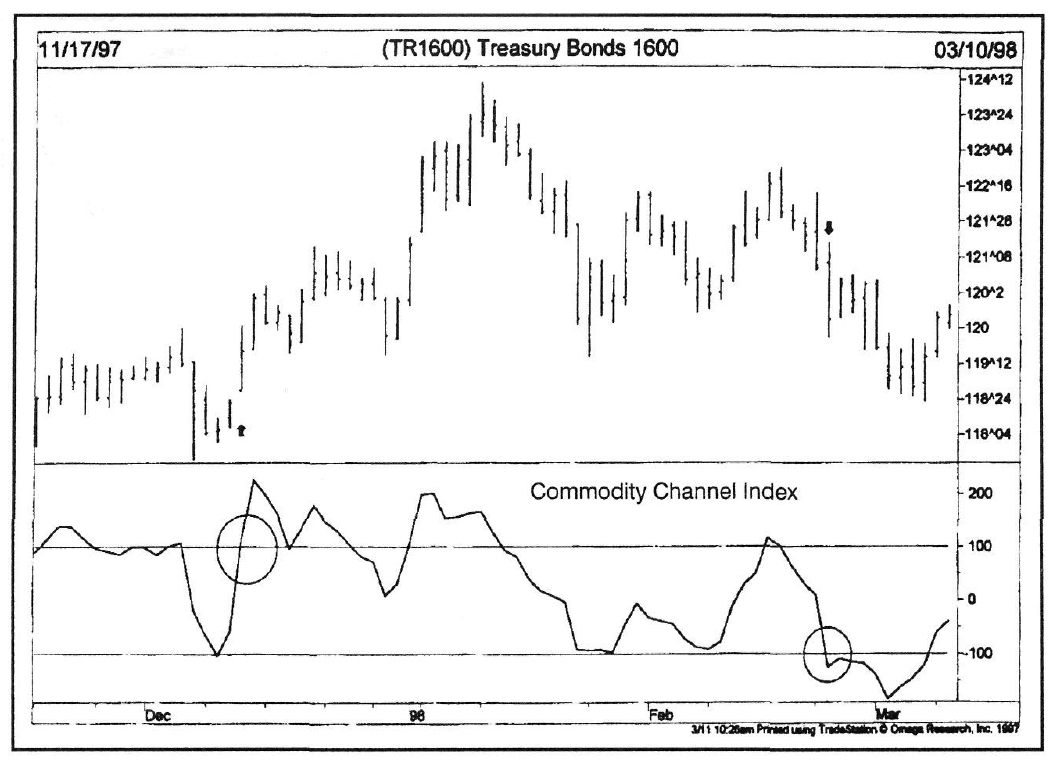
Do inicjalizacji wskaźnika podajemy okres w postaci cyfry, przykładowo 14 to okres domyślny oraz najpopularniejszy i oznacza on że do obliczeń przyjmowany jest okres 14 dni. Do obliczania wartości wskaźnika używany jest następujący wzór:

Do otrzymania średniej wartości wzrostu cen trzeba dodać wszystkie punkty zyskane podczas wzrostów, pod czas przyjętego okresu (w tym przypadku 14 dni) i 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.

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[[17]](#footnote-17).

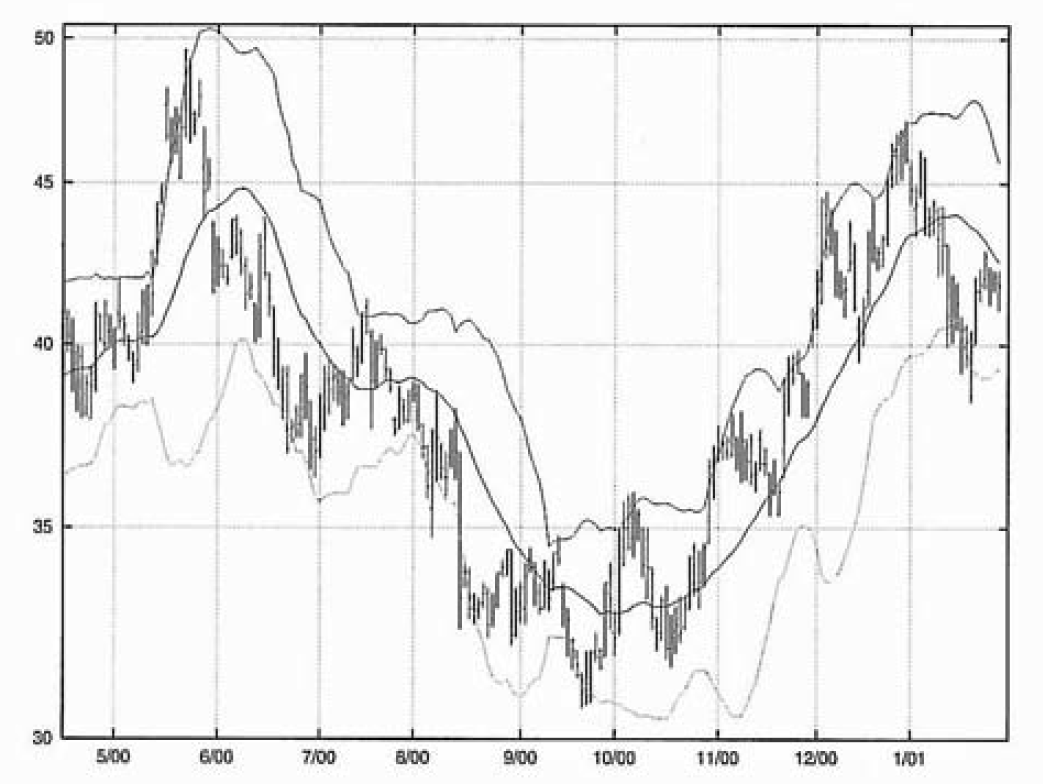
#### CCI - Commodity Channel Index

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[[18]](#footnote-18).



Rysunek 2.12 Wskaźnik CCI  
Źródło: Murphy, Analiza techniczna rynków finansowych, WIG-PRESS Warszawa 1999, s 208

#### 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. 

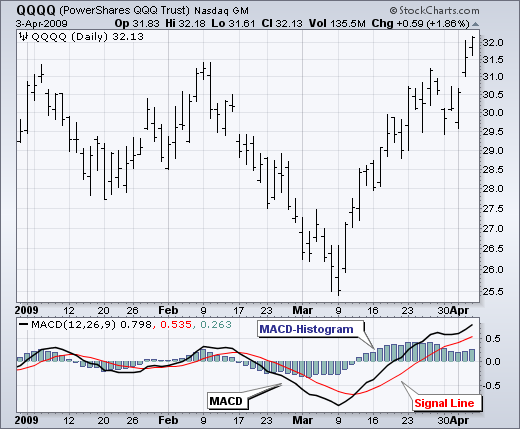
Rysunek 2.13 Bollinger Bands  
Źródło: Bollinger, Bollinger on Bollinger Bands, McGraw-Hill New York 2001, s xxi

Podstawą wstęg Bollingera jest wskaźnik średniej ruchomej, stanowi on również środkową wstęgę opisywanego wskaźnika. Domyślna wartość średniej ruchomej 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 ruchomej. Wstęga górna i dolna rysowane są w odległości dwóch odchyleń standardowych od średniej.

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[[19]](#footnote-19).

#### 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 2.14 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[[20]](#footnote-20).

## 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[[21]](#footnote-21).

### 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[[22]](#footnote-22).

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[[23]](#footnote-23).

### Podstawowe pojęcia

Poniżej zostaną opisane podstawowe pojęcia związane z uczeniem maszynowym użyte w pracy:

* 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 pojedyncza wiersza w zestawie danych, 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,
* 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 w relacji do wszystkich danych z zestawu,
* Skuteczność (ang. Recall) – wartość poprawnie sklasyfikowanych instancji ze wszystkich instancji danej klasy,
* Walidacja krzyżowa – metoda testowania która dzieli 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.

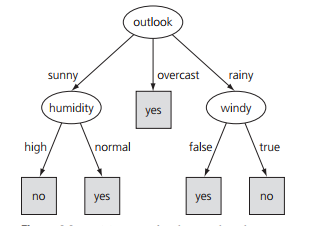
### Przedstawienie badanych algorytmów

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

#### 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 możliwości 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.



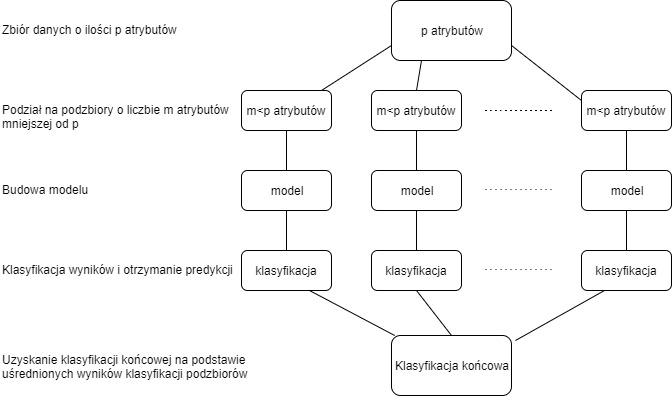
Rysunek 2.15 Przykładowe drzewo  
Źródło: Witten,Eibe,Mark., Data Mining: Practical Machine Learning Tools and Techniques

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.

#### 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 2.16 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[[24]](#footnote-24).

# 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.

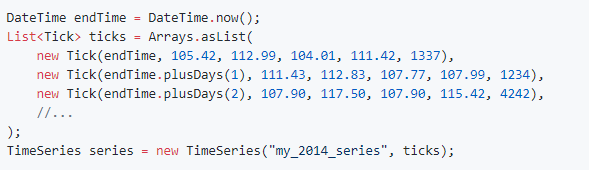
## Technical Analysis for Java (Ta4j)

Ta4j jest to biblioteka napisana przeprowadzania analizy technicznej w języku Java typu open source na licencji MIT. Dostarcza 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, silnik do tworzenia strategii inwestycyjnych.

### 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.

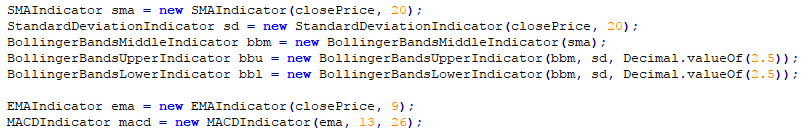
Obiekt Tick jest obiektem zawierającym informacje o pojedynczym zakończonym przedziale czasowym. Jest reprezentacją świecy japońskiej i zawiera informacje o cenie otwarcia, cenie zamknięcia, najwyższej i najniższej cenie oraz wolumenie ceny.



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 oraz dla wskaźników bardziej złożonych wymagane są inne wskaźniki na podstawie których są generowane wartości.



Rysunek 3.2 Wskaźniki  
Źródło: opracowanie własne

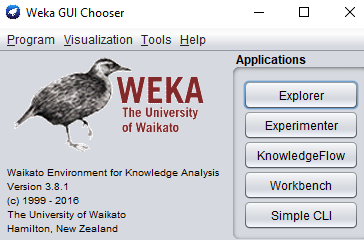
Dodatkowo konstruktory wskaźników przyjmują wartości liczbowe które mogą oznaczać odchylenia czy okresy czasowe jakie przyjmuje wskaźnik. Wszystko zależnie od rodzaju wskaźnika. Wartości wskaźników są obliczane dynamicznie w momencie wywołania metody get (int index) na obiekcie wskaźnika.

Innymi możliwościami biblioteki jest tworzenie strategii, czyli zestawu reguł według którego miały by być podejmowane decyzje kupna/sprzedaży, oraz tester strategii za pomocą którego można przetestować stworzoną strategię na historycznych danych[[25]](#footnote-25).

## Weka

Weka jest to zbiór zawierający 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.



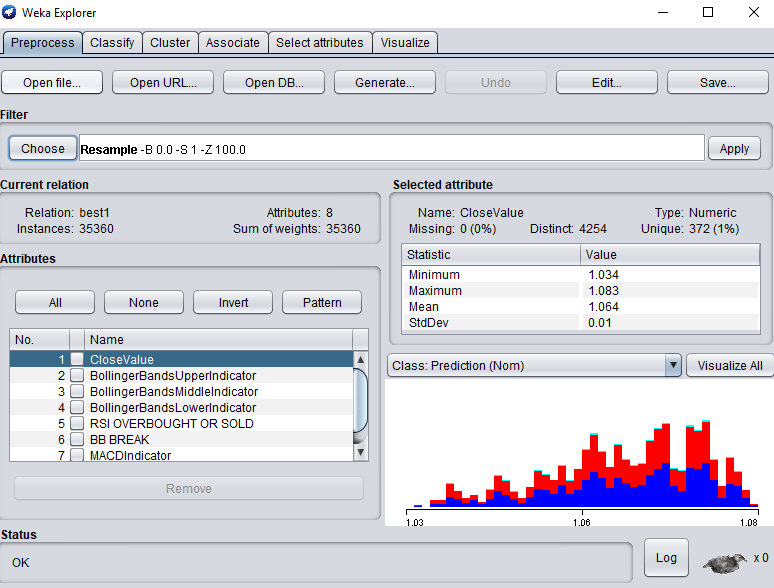
Rysunek 3.3 Weka – aplikacje  
Źródło: opracowanie własne

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

### 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:

* **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.

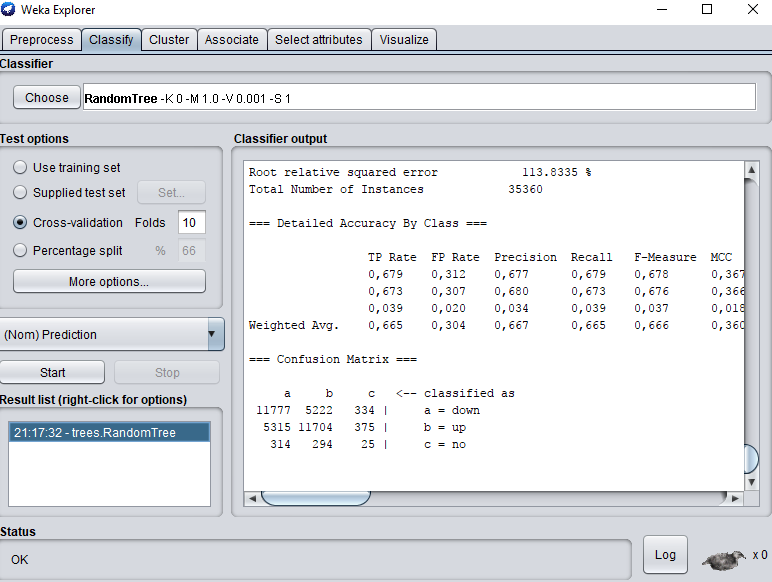


Rysunek 3.4 Sekcja preprocess  
Źródło: opracowanie własne

### 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

### Pozostałe funkcjonalności

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

* **Cluster** – umożliwia wykonanieprocesu 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.

# Metodologia

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

## Cel i przedmiot badań

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

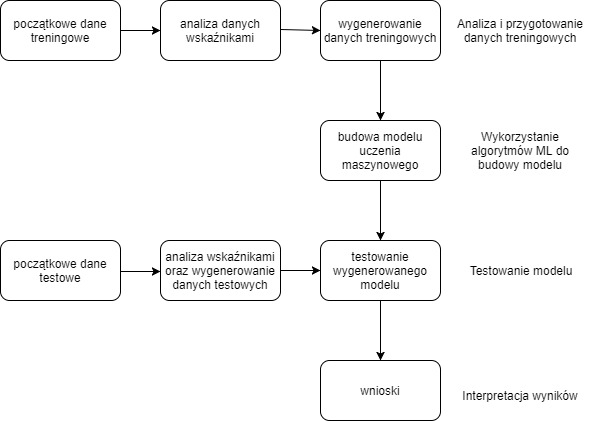
* poszukiwanie optymalnego algorytmu prognozy zmian wybranych par walutowych,
* symulację gry na opcjach binarnych z wykorzystaniem zbudowanego modelu.

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 za określony czas 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, służących jako zbiory wejściowe do algorytmów weki.

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”, skuteczność 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 skuteczność prognozy ruchu w dół określana również jako „Recall Down”.

## 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.

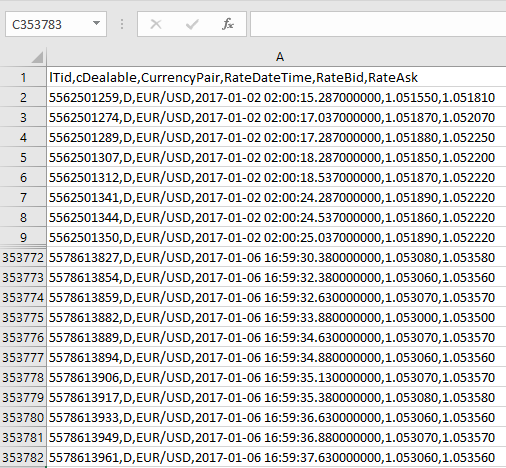


Rysunek 4.1 Etapy pracy  
Źródło: Opracowanie własne

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 wygenerowaniu modelu. 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.

## 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 w zakresie tej pary walut. 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.



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.

## 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.

### 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 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.

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

**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**

### 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.

#### Sygnał RSI

Wskaźnik RSI może generować sygnał prawdopodobnego spadku ceny reprezentowany przez liczbę jeden oraz wzrostu ceny reprezentowany przez liczbę minus jeden. Gdy wartość sygnały wynosi zero oznaczą to informację o braku okazji do decyzji. Zakładając, że wartość poziomu górnego wskaźnika RSI jest oznaczony jako ***x*** a poziom dolny jako ***y*** i wartość wskaźnika RSI to ***rsi*** to generowany sygnał jest opisany następującymi równaniami:

Rysunek 4.4 Równania opisujące sygnał RSI  
Źródło: opracowanie własne

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.

#### 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.

Rysunek 4.5 Równania opisujące sygnał CCI  
Źródło: opracowanie własne

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, w przypadku gdy wartość wskaźnika porusza się miedzy 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.

#### Sygnał Wstęg Bollgingera

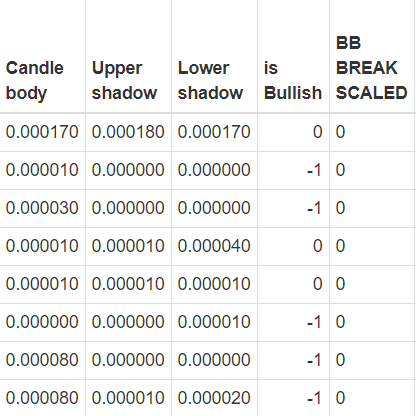
Na podstawie wskaźnika wstęg Bollingera również generowane są sygnały. Reprezentacja podobna jest jak w poprzednich. Generowane są trzy sygnały: jeden, minus jeden, zero. Wskaźnik składa się z linii górnej i dolnej które śledzą ruch ceny. Gdy wartość ceny przekroczy wartość wskaźnika wstęgi górnej generowany 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.

### Wygenerowane dane

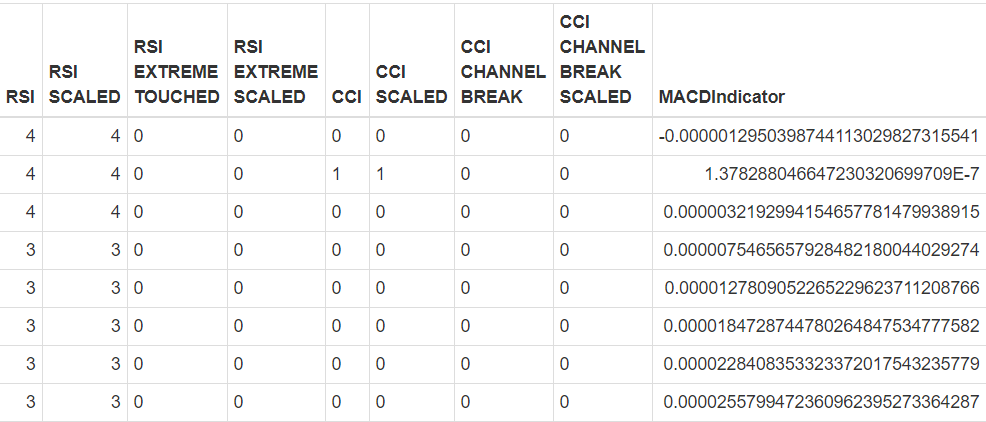
Dane wygenerowane przez aplikację stworzoną na potrzeby badań składają się z wielu elementów takich jak cechy świecy, wartości wskaźników i sygnały. Z miesięcznego zestawu ruchów ceny który zawierał ponad 350 tysięcy wierszy, po zamianie na świece zostało ponad 35 tysięcy wierszy które stanowią dane 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,

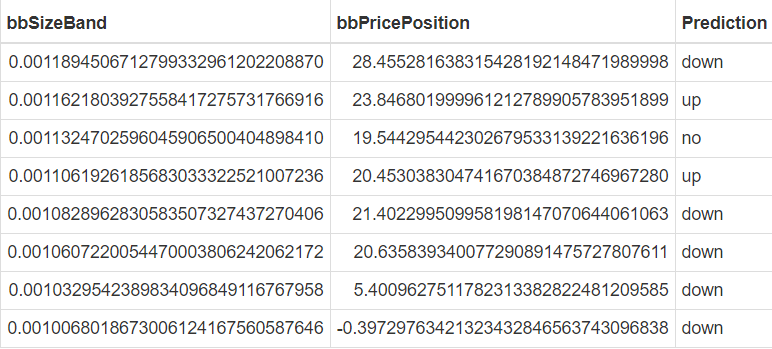


Rysunek 4.6 Wygenerowane dane 1/3  
Źródło: opracowanie własne

* 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, 

Rysunek 4.7 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.



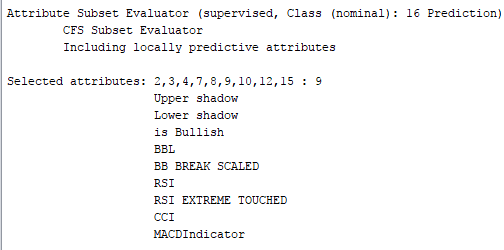
Rysunek 4.8 Wygenerowane dane 2/3  
Źródło: opracowanie własne

## Omówienie procesu badań

Proces badań składał się z czterech poszczególnych etapów. Wyznaczania klasyfikatorów na podstawie całego zbioru atrybutów by wybrać najskuteczniejsze oraz wyznaczania atrybutów, gdzie za pomocą algorytmów oceniających były wyznaczane atrybuty najbardziej znaczące dla badań, kolejno filtrowanie danych, aby pozbyć się atrybutów obniżających wyniki badań, następnie klasyfikacja za pomocą wybranych metod klasyfikujących oraz 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 które najlepiej radzą 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.

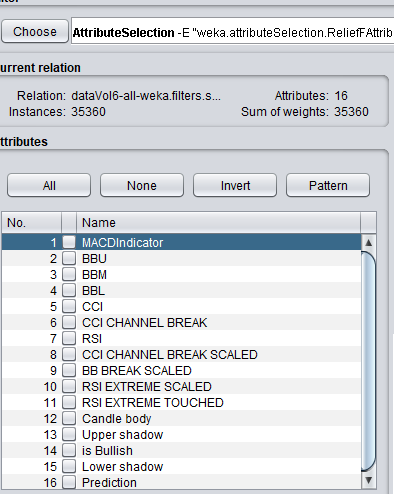
### Wyznaczenie i filtrowanie atrybutów

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.



Rysunek 4.9 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.

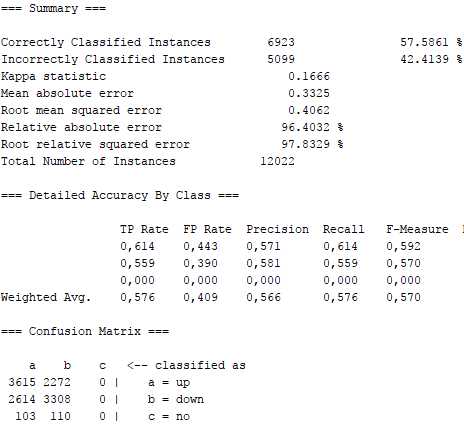


Rysunek 4.10 Zbiór atrybutów po użyciu filtra „AttributeSelection”  
Źródło: opracowanie własne

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.

### 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.



Rysunek 4.11 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 testowy a pozostałe jako zbiór treningowy. W tym etapie celem było uzyskanie modelu o najwyższych parametrach „precision” i „recall” w odniesieniu do wszystkich atrybutów decyzyjnych.

### 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 możliwym 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.

# 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.

## 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:

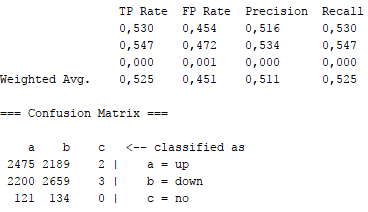
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Algorytm | Precision up | Recall up | Precision down | Recall down | Precision no | Recall no |
| bayes |  |  |  |  |  |  |
| BayesNet | 0,494 | 0,642 | 0,524 | 0,401 | 0 | 0 |
| NaiveBayes | 0,501 | 0,509 | 0,519 | 0,538 | 0 | 0 |
| 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 wybrany jako najlepszy został klasyfikator lasów losowych (ang. Random Forest). Ten klasyfikator zwrócił najlepsze wyniki dla standardowego zbioru danych bez żadnej 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 skuteczność algorytmu, gdzie większa 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 lepiej dużo 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.

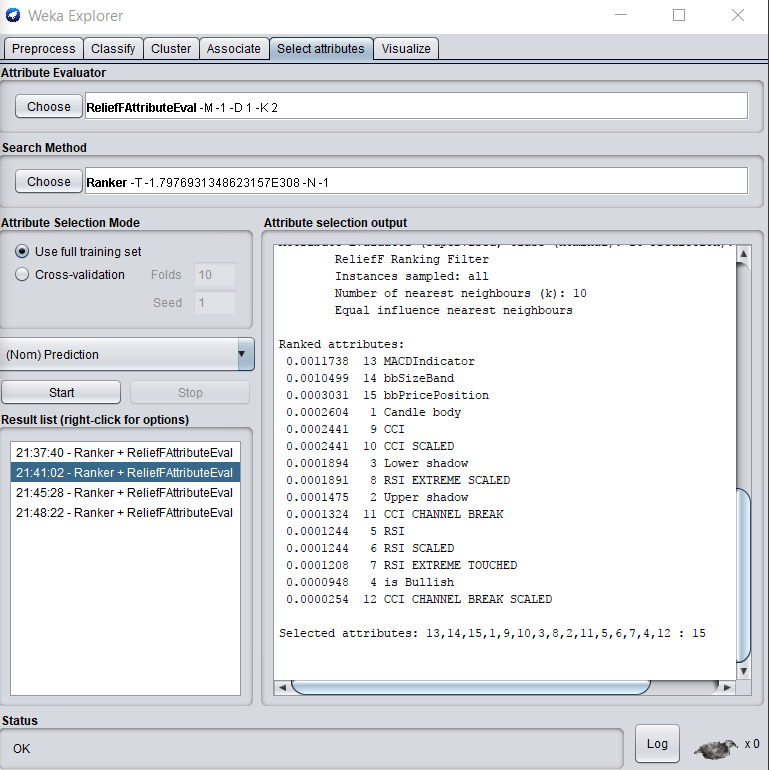
## Atrybuty

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.



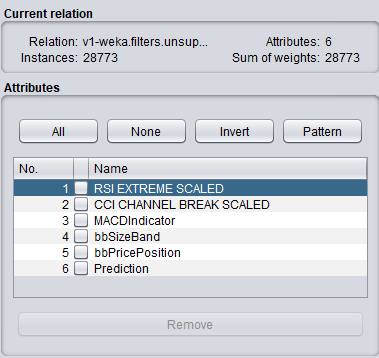
Rysunek 5.1 Skuteczność drzew losowych po selekcji atrybutów  
Źródło: Opracowanie własne

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 1,9,3,10,3 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.

## 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 5.5.2 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 parametrów. Najlepszy model osiągnął w przybliżeniu:

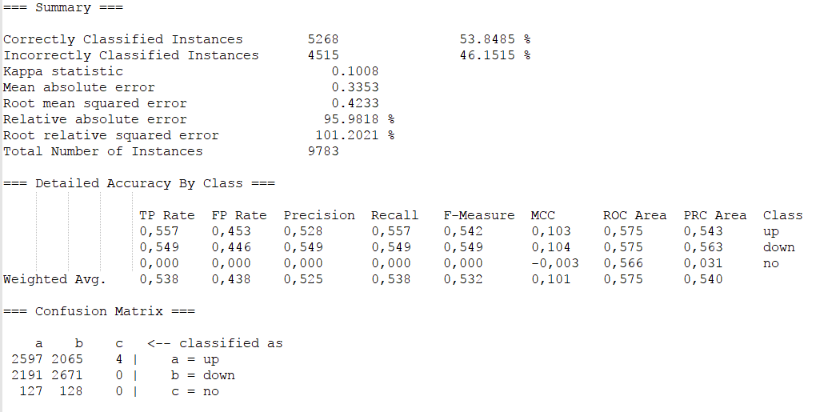
* Precyzję na poziome 52,2% określenia kierunku ruchu w górę, przy 56,1% skuteczności określenia decyzji należącej do prawidłowej klasy decyzyjnej.
* Precyzję na poziome 55,4% określenia kierunku ruchu w dół, przy 55,3% skuteczności określenia decyzji należącej do prawidłowej klasy decyzyjnej.
* Precyzję na poziome 16,4% określenia braku ruchu ceny, przy 4% skuteczności określenia decyzji należącej do prawidłowej klasy decyzyjnej.

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.

## Symulacja - prognoza o pięć minut.

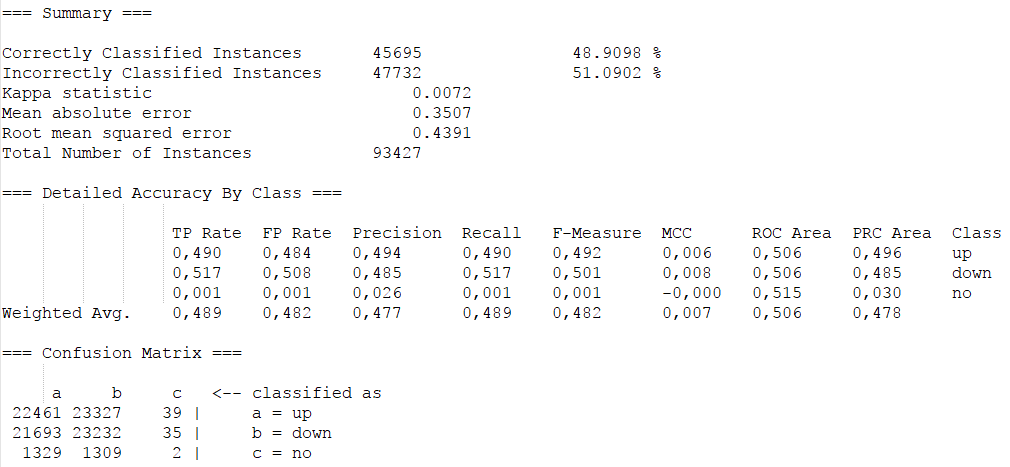
Poniżej na rysunku widać wynik uzyskanego modelu bazującego na parametrach opisanych w poprzednich podrozdziałach który został użyty do przeprowadzenia symulacji przewidywania o pięć minut na podstawie danych testowych z trzech miesięcy.



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

Precyzja uzyskanego modelu wynosi 52.8% przy przewidywaniu ruchów w górę i 54.9% przy ruchach w dół. Skuteczność podjętych decyzji to 55.7% przy ruchach w górę i 54.9% przy ruchach w dół. Sklasyfikowane zostało poprawnie 53.8% instancji.

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



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 skuteczność predykcji braku ruchu jest bardzo mała ze względu na to, że brak ruchu na ruchu giełdowym 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 skuteczność dla każdego z testowanych miesięcy. Widać nieznaczne różnice w precyzji predykcji kierunku w dół oraz w skutecznościach 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 skuteczności określania instancji. Dla ruchów w górę skuteczność mieści się w zakresie od 46% do 51%, natomiast dla ruchów w dół od 52% do 55%.

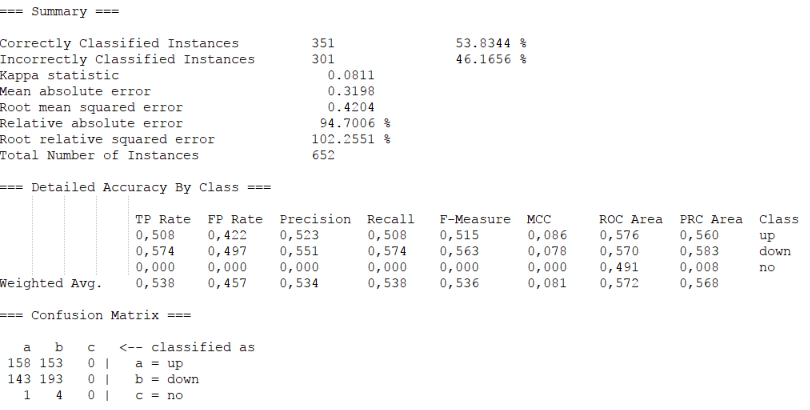
Rysunek 5.6 Wykres wyników precyzji predykcji o 5 min z podziałem na dni  
Źródło: opracowanie własne

Rysunek 5.7 Wykres wyników skuteczności 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 skuteczności, 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 skuteczność w ocenie ruchów w dół. Anomalie w postaci pojedynczych wzrostów i spadków predykcji oraz skuteczności 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.

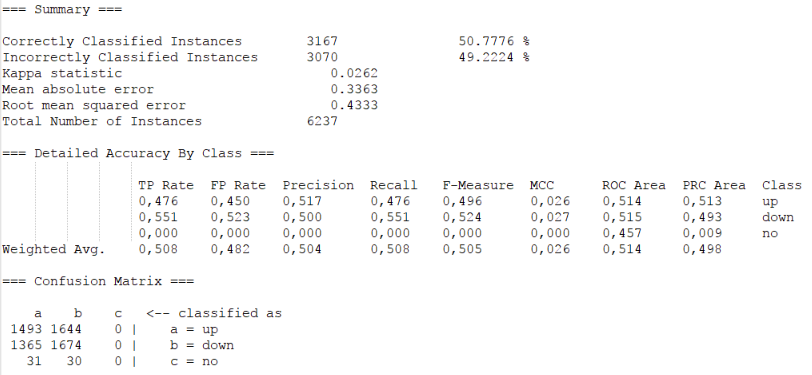
## Symulacja - prognoza trzydzieści minut

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.



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. Skuteczność modelu wynosi 50.8% dla ruchów w górę i 57.4% dla ruchów w dół. Wartość poprawnie sklasyfikowanych instancji wynosi 53.8%.



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 lekki wzrost względem wyników tego badania na danych predykcji o pięć minut. Precyzja określenia ruchów w górę wynosi 51.7%, a ruchów w dół 50%. Skuteczność 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 skuteczności 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 skuteczności dla ruchów w górę wynosi od 41% do 55%, a dla ruchów w dół od 51% do 61%.

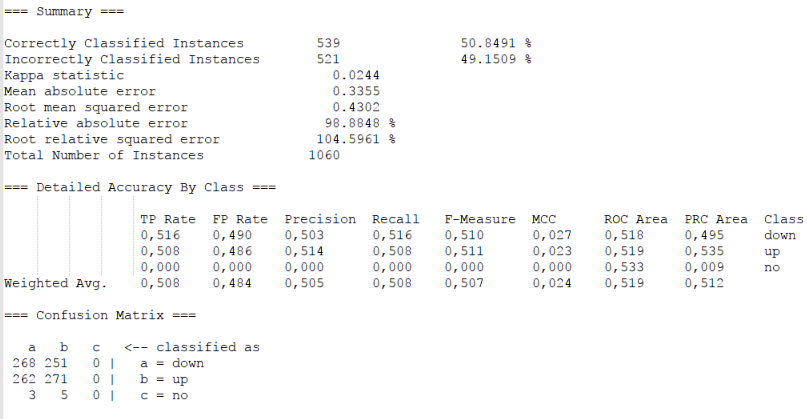
Rysunek 5.100 Wykres wyników precyzji predykcji o 30 min z podziałem na dni  
Źródło: opracowanie własne

Rysunek 5.11 Wykres wyników skuteczności 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 45% do 57%. Skuteczności 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.

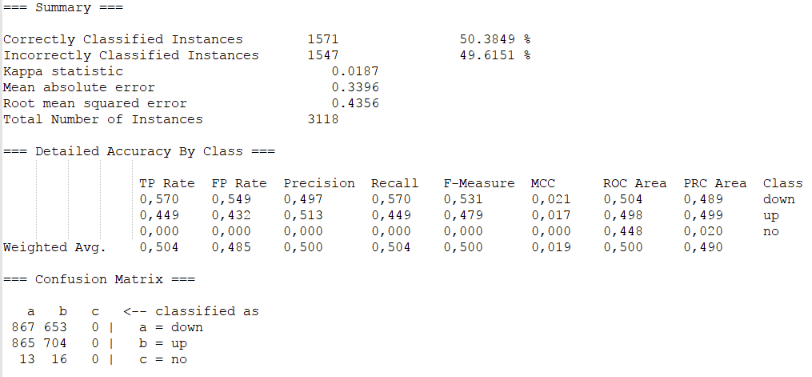
## Symulacja - prognoza o godzinę

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.



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 skuteczności równej 50.8%, a precyzja predykcji ruchów w dół wyniosła 50.3% przy skuteczności 51.6%. Sklasyfikowanych prawidłowo zostało 49.1% obiektów.



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 skuteczności względem tej w podsumowaniu modelu. Dla przewidywania ruchów w górę precyzja wyniosła 51.3%, a skuteczność 44.9%. Dla przewidywania ruchów w dół precyzja wyniosła 49.7%, a skuteczność 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%. Skuteczność 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.

## Podsumowanie symulacji

Wynikiem badań jest wybrany optymalny algorytm uczenia maszynowego dający najlepsze rezultaty z badanych algorytmów. Dla wszystkich symulacji wyniki wyszły bardzo zbliżone do siebie. Precyzja i skuteczność prognozy ruchów w każdym z przypadków oscylowała w okolicach 50%. Zróżnicowanie ze wyników przy podziale na miesiące i tygodnie nie było duże, ponieważ sięgało nie więcej niż 4% w przypadku precyzji i 6% w przypadku skuteczności 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 i częstymi wydarzeniami są niespodziewane ruchy. Dodatkowo rozbicie zbioru danych na mniejsze części sprzyjało większym wahaniom wyników, ponieważ im mniejszy zbiór tym większy 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 algorytmu 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.

# 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 skuteczności wąchający się w okolicach 50% jest wynikiem na średnim poziomie, jednak połowa instancji była klasyfikowana prawidłowo przy tak losowym i zawierającym wiele hałasów 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 hałasów 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 hałaśliwe.

Obszar dotyczący wykrywania zależności w tego typu dużych zbiorach hałaśliwych 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 jednak jest to tak obszerny temat, że nadawał by się na kolejną pracę magisterską. Warto też ograniczyć liczbę podejmowanych decyzji do tych w czasie, gdy cena znajduje się w ekstremalnych wartościach wskaźników co prawdopodobnie spowodowałoby spadek hałasów.

# Bibliografia

WEKA Manual for Version 3-9-1

<http://www.cs.waikato.ac.nz/ml/weka/index.html>

<https://github.com/mdeverdelhan/ta4j>

[https://nlp.stanford.edu/IR-book/html/htmledition/text-classification-and-naive-bayes-1.html 24.07.2017](https://nlp.stanford.edu/IR-book/html/htmledition/text-classification-and-naive-bayes-1.html%2024.07.2017)

https://devstyle.pl/2017/08/09/machine-learning-w-ceneo/

1. J. Mazurek http://www.bankier.pl/wiadomosc/Czym-jest-rynek-Forex-1797860.html (stan na 27.06.2017) [↑](#footnote-ref-1)
2. http://www.sharptrader.com/new-to-trading/forex/majors-minors-exotic-currency-pairs/ (stan na 27.06.2017) [↑](#footnote-ref-2)
3. M.C.Thomsett, *Mastering Fundametnal Analysys,* Dearborn Financial Publishing 1998, s 2-3 [↑](#footnote-ref-3)
4. J.J.Murphy, *Analiza techniczna rynków finansowych, WIG-PRESS Warszawa 1999, s 1* [↑](#footnote-ref-4)
5. 9 10 .J.Murphy, *Analiza techniczna rynków finansowych, WIG-PRESS Warszawa 1999, s 2* [↑](#footnote-ref-5)
6. [↑](#footnote-ref-6)
7. http://www.investopedia.com/university/technical/techanalysis1.asp (stan na 04.06.2017) [↑](#footnote-ref-7)
8. S.Nison, *Japanise candlestick charing techniques*, Prentice Hall Press USA 2001, s 16 [↑](#footnote-ref-8)
9. S.Nison, Japanise candlestick charing techniques, Prentice Hall Press USA 2001, s 21-26 [↑](#footnote-ref-9)
10. S.Nison, Japanise candlestick charing techniques, Prentice Hall Press USA 2001, s 149-150 [↑](#footnote-ref-10)
11. .J.Murphy, *Analiza techniczna rynków finansowych, WIG-PRESS Warszawa 1999, s 42* [↑](#footnote-ref-11)
12. .J.Murphy, *Analiza techniczna rynków finansowych, WIG-PRESS Warszawa 1999, s 42-47* [↑](#footnote-ref-12)
13. http://www.investopedia.com/university/technical/techanalysis10.asp (dostęp 10.07.2017) [↑](#footnote-ref-13)
14. J.Murphy, *Analiza techniczna rynków finansowych, WIG-PRESS Warszawa 1999, s 197-198* [↑](#footnote-ref-14)
15. http://www.investopedia.com/terms/c/crossover.asp (dostęp 11.07.2017) [↑](#footnote-ref-15)
16. J.Murphy, *Analiza techniczna rynków finansowych, WIG-PRESS Warszawa 1999, s 137* [↑](#footnote-ref-16)
17. J.Murphy, *Analiza techniczna rynków finansowych, WIG-PRESS Warszawa 1999, s*  [↑](#footnote-ref-17)
18. <https://www.linnsoft.com/techind/cci-commodity-channel-index> (dostęp na 14.07.2017) [↑](#footnote-ref-18)
19. J. Bollinger, *Bollinger on Bollinger Bands,* McGraw-HillNew York 2001, s xx-xxii [↑](#footnote-ref-19)
20. http://stockcharts.com/school/doku.php?id=chart\_school:technical\_indicators:moving\_average\_

    convergence\_divergence\_macd (dostęp na 17.07.2017) [↑](#footnote-ref-20)
21. S. Shalev-Shwartz, S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms,*

    Cambridge University Press 2014, s vii [↑](#footnote-ref-21)
22. P.Cichosz, *Systemy uczące się*, Wydawnictwa Naukowo-Techniczne, Warszawa, 2000 [↑](#footnote-ref-22)
23. S. Shalev-Shwartz, S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms,*

    Cambridge University Press 2014, s 19-22 [↑](#footnote-ref-23)
24. http://isplab.tudelft.nl/sites/default/files/Lai05a.pdf (dostęp na 26.07.2017) [↑](#footnote-ref-24)
25. https://github.com/mdeverdelhan/ta4j (dostęp na 25.07.2017) [↑](#footnote-ref-25)