Zastosowanie metody drzewa decyzyjnego w analizie problemów makroekonomicznych

Application of the decision tree method in macroeconomics problems analysis

Marta Zalewska

Szkoła Główna Handlowa w Warszawie

Wojciech Zalewski

Politechnika Białostocka, Wydział Zarządzania, Katedra Informatyki Gospodarczej i Logistyki

Abstract

Decision tree is a method of data mining analysis based on extensive set of independent variables. Main advantage of this approach is that not only does it provide an interesting visualization of the problem, but also creates a model of a very good quality. This paper presents the decision tree method in an multi-criteria analysis of determinants of foreign direct investment inflows. The analysis is based on data from the World Bank and Doing Business reports for the years 2006-2010.

***Keywords**:* decision making, decision tree, foreign direct investment, data mining

Wstęp

Wykorzystanie data mining jest dobrą metodą analizy dużych zbiorów danych przy skromnym zestawie warunków początkowych. W celu wyodrębnienia szerokiej gamy powiązań unika się ustalania ograniczeń *a priori*. Artykuł ma na celu prezentację metody J48, która używa algorytmu C4.5 do wygenerowania drzewa klasyfikującego.

Zastosowanie metody drzewa J48 zaprezentowano na przykładzie analizy determinant napływu bezpośrednich inwestycji zagranicznych. Analiza ilościowa opiera się na danych uzyskanych z Raportu *Doing Business* oraz z Banku Światowego. Są to dane roczne z lat 2006-2010 na poziomie kraju. Zawierają zarówno informacje o stanie gospodarki takie jak PKB lub populacja, ale również parametry swobody prowadzenia działalności gospodarczej w danym kraju. Zmienną objaśnianą jest wysokość FDI (ang. *Foreign Direct Investment* – bezpośrednie inwestycje zagraniczne).

Na podstawie uzyskanych danych wygenerowano drzewo decyzyjne przy użyciu oprogramowania Weka udostępnianego przez Uniwersytet w Waikato w Nowej Zelandii. Otrzymany model charakteryzuje się współczynnikiem Kappa powyżej 60%, a macierz pomyłek wskazuje na dużą liczbę poprawnie sklasyfikowanych instancji, co oznacza bardzo dobrą jakość drzewa.

1. Dotychczasowe metody badania FDI

Dotychczasowe badania bezpośrednich inwestycji zagranicznych wykorzystywały wnioski z obszaru teorii makroekonomii. W tym ujęciu zidentyfikowano takie determinanty FDI jak polityka fiskalna państwa, poziom zaufania do obcej gospodarki oraz różnice w funkcjonowaniu rynków kapitałowych krajów[1]. Większość badań w literaturze przedmiotu ma charakter jedynie rozważań teoretycznych, opartych głównie o teorię makroekonomii, część zaś stanowi jedynie powierzchowną analizę tego wielowymiarowego zjawiska, opartą na wąskim zestawie zmiennych objaśniających. Aby wypełnić tę lukę, zebrane dane poddano analizie ilościowej metodą uczenia się maszynowego (*machine learning*).

W programie Weka wygenerowano drzewo klasyfikacyjne. Taki wybór metody podyktowany został względami praktycznymi – drzewo decyzyjne stanowi ciekawą wizualizację zaobserwowanych zależności przy zadowalającej jakości modelu.

[1] Mankiw G., 2009. *Principles of Macroeconomics.* South-Western Cengage Learning, s. 389

2. Algorytm C4.5

Drzewo J48 jest generowane przy użyciu algorytmu C4.5[2], który dzieli pierwotny zestaw danych względem każdej ze zmiennych. W ten sposób powstaje tyle wariantów podziału, ile w zestawie jest zmiennych objaśniających. Dla każdego podziału liczona jest wartość metryki *information gain*, która zdefiniowana jest jako przyrost entropii w każdym z podzbiorów. Zmienna o najwyższym współczynniku *information gain* staje się pierwszym węzłem drzewa. Następnie dla wszystkich podzbiorów powtarza się tę operację aż do wyczerpania wszystkich instancji. Przebieg procesu można przedstawić w następujących krokach:

- należy podzielić zestaw danych *E* według każdej zmiennej i policzyć wartość *information gain*, czyli przyrostu entropii uzyskanych podzbiorów w stosunku do zbioru pierwotnego według wzoru $p[\log(p/t)-\log(P/T)]$;
- należy wybrać zmienną *a*, zapewniającą najwyższy przyrost informacji i według niej podzielić zbiór pierwotny;
- na każdym podzbiorze należy powtarzać operację aż do wyczerpania instancji.

3. Mierniki jakości modelu

W metodyce *data mining* najczęściej wykorzystuje się następujące mierniki: *TP Rate*, *FP Rate*, *Precision*, *Recall*, *F-measure* oraz statystykę *Kappa*. Miara *TP Rate* pokazuje, jaki odsetek obserwacji z danej klasy jest poprawnie klasyfikowany przez model, czyli liczy przypadki *true positive*. *FP Rate* opisuje, jaka część obserwacji nienależących do danej klasy to obserwacje błędnie do niej zaklasyfikowane – *false positive*. *Precision* to miernik precyzji przyporządkowania danej obserwacji do adekwatnej klasy. Kategoria *Recall* pokazuje poprawne pokrycie danej klasy. Miara *F-measure* to ogólny wskaźnik jakości, który wylicza się na podstawie wzoru:

$$F-measure=\frac{2\cdot recall\cdot precision}{recall+precision}$$

Statystyka *Kappa* mierzy zgodność między proponowanym przydziałem instancji do klasy a stanem faktycznym, co stanowi o ogólnej trafności modelu. Kształt krzywej ROC pokazuje jak wygenerowany model tłumaczy rzeczywistość – im bardziej wygięta jest krzywa, tym model lepszy jest od składnika losowego.

[2] Witten I. H., Eibe F., 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

4. Źródło danych wykorzystanych w modelu

Do budowy modelu wykorzystano dane zawarte w raporcie *Doing Business*, tworzonym corocznie przez Bank Światowy. Opracowanie ma na celu ocenę regulacji i ograniczeń, na jakie napotykają prowadzący działalność gospodarczą w danym kraju. Raport ma postać zestawienia tabelarycznego, w którym każdemu państwu przyporządkowane są wartości odpowiednich mierników. Badanie swym zasięgiem obejmuje 183 kraje świata. Na potrzeby budowy modelu dane wzbogacono o dodatkowe informacje pochodzące również z Banku Światowego, który z wyjątkową dokładnością zbiera dane ekonomiczne i społeczne w wielu krajach, z uwzględnieniem tych najbardziej biednych[3]. Wybór źródła danych był podyktowany renomą Banku Światowego oraz ogólnie uznaną rzetelnością jego badań. Z kolei bazy danych raportu *Doing Business* charakteryzują się ukierunkowanymi i przejrzyście opisanymi badaniami, co jest dodatkowym walorem.

Zbiór danych wykorzystany w analizie liczy 901 obserwacji. Jego zakres czasowy to 5 lat (od 2006 do 2010 roku). Są to dane roczne dla 183 państw świata, pogrupowane w 37 kategorii zmiennych objaśniających. Za zmienną objaśnianą w badaniu przyjęto FDI – napływy zagranicznych inwestycji bezpośrednich netto w danym kraju, w dolarach amerykańskich w wartości bieżącej. Zmienną objaśnianą przekształcono do postaci nominalnej, dokonując dyskretyzacji za pomocą nienadzorowanego filtru *Discretize* z opcją równej częstości w każdym z trzech przedziałów, którym nadano etykiety „niskie", „średnie" i „wysokie".

Jako atrybutów objaśniających w modelu użyto zmiennych z raportu *Doing Business* z lat 2006-2010, które przedstawiono w Tabeli 1.

Tabela 1. Zmienne objaśniające pozyskane z raportu Doing Business

Kategoria	Zmienne objaśniające
Zakładanie firmy (starting a business)	o procedury (*procedures1*) – zdefiniowane jako interakcje między założycielami firmy a jednostkami zewnętrznymi, ilość procedur; o czas (*time1*) – mierzony w dniach potrzebny na zrealizowanie danej procedury; o koszt (*cost1*) – mierzony jako procent zysku per capita koszt wynikający z obowiązkowych płatności przedsiębiorstwa, regulowany prawnie; o minimalny wkład własny (*paid in minimum capital*) – kapitał zdeponowany przez przedsiębiorcę przed zarejestrowaniem firmy, procent zysku per capita.

[3] Strona Banku Światowego. Tryb dostępu: http://www.worldbank.org, stan z dn. 01.06.2012.

cd. Tabeli 1.

Uzyskiwanie pozwoleń na budowę (dealing with construction permits)	o procedury (*procedures2*) – ilość procedur opisanych jak wcześniej; o czas (*time2*) – czas zrealizowania danej procedury mierzony w dniach; o koszt (*cost2*) – procent wartości danego dobra, zidentyfikowany jak wcześniej.
Rejestrowanie własności (registering property) – proces kupna nieruchomości i transferu praw własności ze sprzedającego na nabywcę	o procedury (*procedures3*) – jak wcześniej, o czas (*time3*) – jak wcześniej, o koszt (*cost3*) – mierzony jako procent wartości danego dobra
Otrzymywanie kredytu (getting credit) – opisuje łatwość w otrzymaniu kredytu wynikającą z przepisów oraz łatwość dostępu do informacji o kredycie	o siła regulacji prawnych (*strength of legal rights index*) – wartości 0-10, im wyższa, tym łatwiej jest otrzymać kredyt; mierzy, jak przepisy chronią prawa pożyczkodawców i jak wpływa to na łatwość dostępu do kredytu, o dostęp do informacji o kredycie (*depth of credit information index*) – wartości 0-6, im większa wartość, tym lepszy dostęp do informacji o kredycie, o publiczny rejestr kredytów (*public credit registry coverage*) – ilość osób i firm, których historia kredytowa z ostatnich 5 lat jest zarejestrowana w rejestrze publicznym, wyrażony jako procent z populacji dorosłych, o prywatny rejestr kredytów (*private credit bureau coverage*) – ilość osób i firm, których historia kredytowa z ostatnich 5 lat jest zarejestrowana w rejestrach prywatnych, wskaźnik wyrażony jako procent z populacji dorosłych
Ochrona inwestorów (protecting investors) – mierzy siłę ochrony drobnych inwestorów i udziałowców przed przejęciem firmowych środków dla osobistej korzyści przez kadry kierownicze	o poziom ujawnienia (*extend of disclosure index*) –wartości 0-10, mierzy stopień wyjawiania szczegółów transakcji w firmie, o poziom odpowiedzialności dyrektora (*extend of director liability index*) –wartości 0-10, wyższe oznaczają większą odpowiedzialność dyrektora za podejmowane przez firmę decyzje, o siła udziałowców (*ease of shareholder suits index*) – wartości 0-10, wyższy oznacza większą siłę udziałowców do kwestionowania transakcji, o poziom ochrony inwestorów (*strength of invesor protection index*) – stanowi średnią arytmetyczną powyższych wskaźników i mierzy ogólnie pojęty poziom ochrony inwestorów, wartości 0-10

cd. Tabeli 1.

Płacenie podatków (paying taxes) – dotyczy podatków i obowiązkowych opłat, które firma średniej wielkości musi uiścić w ciągu roku	o płatności (*payments*) – odzwierciedla ilość wszystkich podatków i opłat, które przedsiębiorstwo musi uiścić w ciągu roku, zmienna wyrażona jako ilość opłat, o czas (*time6*) –wyrażony w godzinach w roku, mierzy czas potrzebny do przygotowania i zapłaty podatku od zysków, VAT-u oraz składek społecznych, o ogólna stopa podatkowa (*total tax rate*) – wskaźnik wyrażający procent wszystkich podatków i opłat w odniesieniu do całego rocznego zysku firmy
Handel zagraniczny (trading across borders) – dotyczy importu i eksportu ładunku drogą morską	o dokumenty do eksportu (*documents to eksport*) – ilość dokumentów potrzebnych do eksportu o czas eksportu (*time to eksport*) – czas potrzebny na eksport jednego ładunku, wyrażony w dniach o koszt eksportu (*cost to eksport*) – zawiera opłaty związane z eksportem oraz opłaty za transport, wyrażony w dolarach za kontener o dokumenty do importu (*documents to import*) – ilość dokumentów potrzebnych do importu o czas importu (*time to import*) – czas potrzebny na import jednego ładunku, w dniach o koszt importu (*cost to import*) – zawiera opłaty związane z importem oraz za transport, wyrażone w dolarach za kontener
Zawieranie umów (enforcing contracts) – wskaźniki mierzące skuteczność systemu sądowego w dziedzinie handlowych sporów	o procedury (*procedures8*) – ilość zrealizowanych procedur potrzebna przed przedstawieniem sprawy przed sądem o czas (*time8*) – czas liczony od momentu wniesienia pozwu do zakończenia sprawy, mierzony w dniach o koszt (*cost8*) – liczony jako procent wartości przedmiotu sprawy
Likwidacja przedsiębiorstwa (closing a business) – dotyczy procesów związanych z niewypłacalnością i zakończeniem działalności gospodarczej	o czas (*time9*) – czas potrzebny na odzyskanie kredytu przez pożyczkodawcę wyrażony w dniach o koszt (*cost9*) – koszt procesu - procent wartości własności dłużnika o stopa odzysku (*recovery rate*) – wskaźnik liczony jako centy z dolara odzyskane przez pożyczkodawców przez restrukturyzację, upłynnienie majątku lub odzyskanie długu

Źródło: opracowanie własne.

Dodatkowo, w celu rozszerzenia analizy, do zbioru dołączono statystyki z Banku Światowego opisujące gospodarki:

- liczba ludności (*population*) – liczba osób przebywających w kraju oprócz uchodźców niezarejestrowanych w danym kraju,
- tempo wzrostu PKB (*GDP growth*) – roczny procentowy wzrost PKB kraju,
- PKB per capita (*GDP per capita*) – zmienna wyrażona w dolarach.

Tak skompletowany zestaw danych wykorzystano w dalszej analizie.

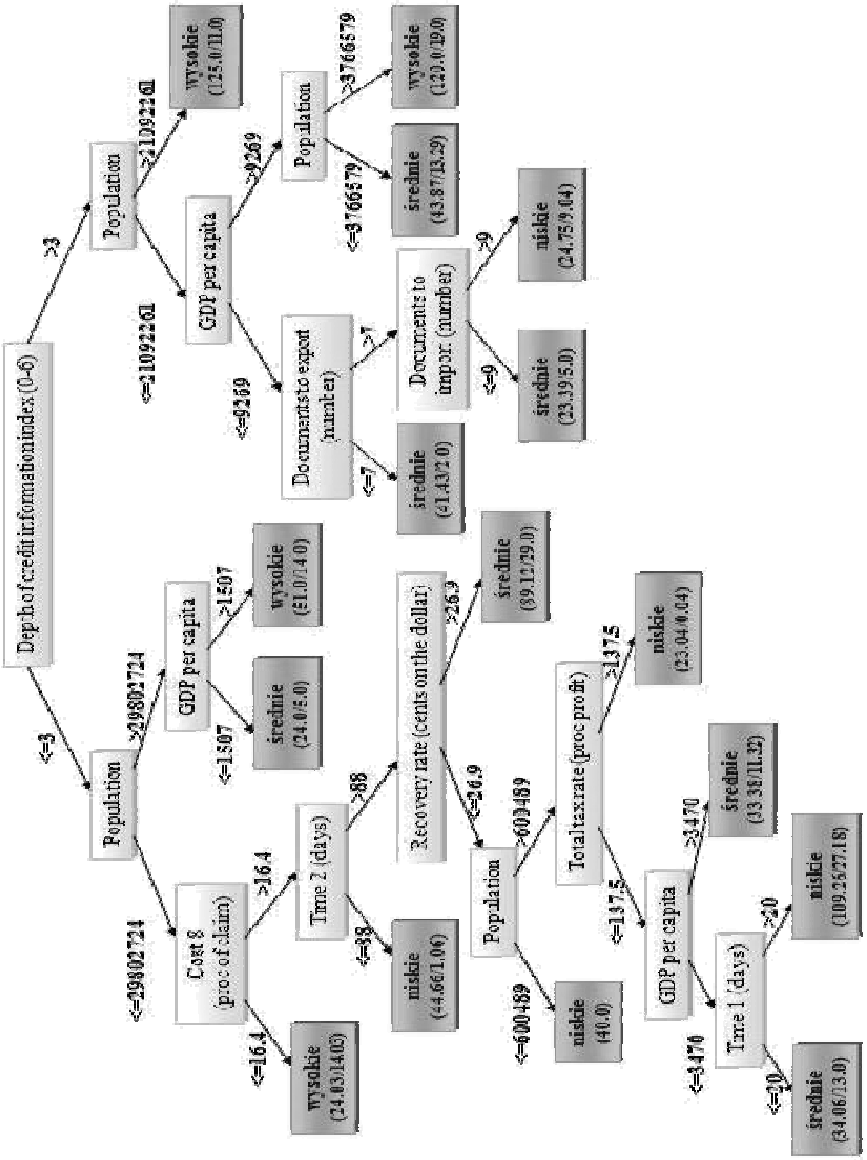
5. Postać wygenerowanego drzewa J48

Wygenerowane drzewo decyzyjne J48 ma głębokość równą 8 poziomów oraz w sumie 16 liści. W korzeniu drzewa znajduje się atrybut „*Credit information index*". Test przeprowadzony na tym atrybucie dzieli zbiór na dwa podzbiory: państw w których indeks ten przyjmuje wartość wyższą od 3 oraz niższą bądź równą 3. Otrzymane w ten sposób poddrzewa są niesymetryczne. Poddrzewo, do którego trafiły państwa z niższą wartością indeksu jest głębsze i bardziej rozczłonkowane. W sumie 4 liście drzewa dotyczą klasy „wysokie", 7 liści wskazuje na klasę „średnie", natomiast 5 na „niskie". Pełną postać otrzymanego drzewa przedstawiono na rys. 1. W Tabeli 2. przedstawiono zestaw mierników jakości modelu J48. Dla każdej klasy ilość instancji *true positive* przekracza 60%, co jest zadowalającym wynikiem. Stopa *false positive* może być uznana za niską, co także świadczy pozytywnie o jakości wygenerowanych modeli.

Tabela 2. Mierniki jakości modelu J48

Klasa	TP Rate	FP Rate	Precision	Recall	F-Measure
wysokie	0,866	0,106	0,803	0,866	0,833
średnie	0,634	0,203	0,610	0,634	0,622
niskie	0,658	0,11	0,745	0,658	0,699

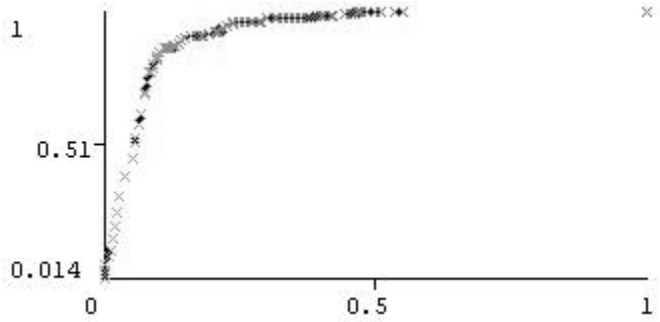
Źródło: opracowanie własne.



Źródło: opracowanie własne.

Rys. 1. Drzewo decyzyjne wygenerowane w programie Weka 6.4 z wykorzystanie algorytmu J48

Dobrą jakość wygenerowanego drzewa potwierdza krzywa ROC (*Receiver Operating Characteristic*) przedstawiona na rys 2. Warto zwrócić uwagę na jej pożądany wygląd – wyraźna wklęsłość, która świadczy o dużej trafności klasyfikacji i wskazuje, iż model jest dużo lepszy niż klasyfikator losowy.



Źródło: opracowanie własne.

Rys. 2. Krzywa ROC dla drzewa J48 przy wartości klasy równej „wysokie FDI"

Pierwszym etapem oceny jakości modelu jest analiza jego kluczowych statystyk. Pozwala to na wstępne określenie trafności wygenerowanych reguł decyzyjnych oraz dostarcza informacji na temat rodzaju i struktury ewentualnych błędów. Tabela 3. zawiera zestawienie najistotniejszych wielkości dla uzyskanego drzewa decyzyjnego.

Tabela 3. Statystyki dotyczące modelu J48

Czynnik	**Wartość**
Statystyka *Kappa*	0,5787
Poprawnie sklasyfikowane instancje	71,91%
Błędnie sklasyfikowane instancje	28.08%
MAE	0.2402
RMSE	0.3646

Źródło: opracowanie własne.

Jak widać statystyka *Kappa* jest bliska 60%. Fakt ten świadczy o tym, że otrzymany model jest o blisko 2/3 lepszy niż klasyfikator losowy, czyli że poprawnie klasyfikuje niemal 60% obserwacji, z którymi klasyfikator losowy sobie nie poradził. Odsetek instancji poprawnie sklasyfikowanych przez drzewo decyzyjne przekracza 70%. Średni błąd absolutny (MAE) modeli wynosi 0,24. Pierwiastek

błędu średniokwadratowego ma wartość 0,36. W Tabeli 4. przedstawiono macierz kontyngencji dla modelu J48.

Tabela 4. Macierz kontyngencji dla modelu J48

wysokie	**średnie**	**niskie**	**J48**
245	36	2	**wysokie**
42	180	62	**średnie**
18	79	187	**niskie**

Źródło: opracowanie własne.

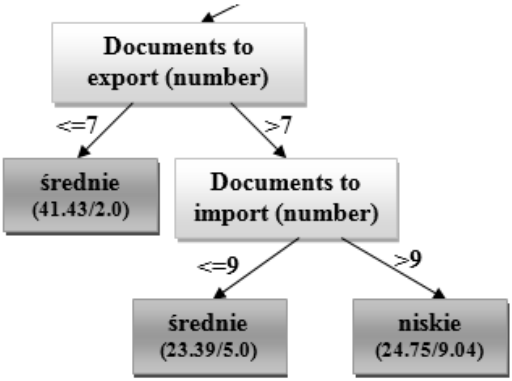
W wierszach znajdują się wartości prognozowane, w kolumnach zaś empiryczne. Na uwagę zasługują wartości elementów na głównej przekątnej, które znacznie przewyższają wartości pozostałych komórek, oznaczających poprawnie sklasyfikowane instancje, co przemawia na rzecz dużej trafności drzewa decyzyjnego.

6. Interpretacja wyników

Zastosowanie metody drzewa decyzyjnego ujawniło różny wpływ zmiennych objaśniających na bezpośrednie inwestycje zagraniczne (FDI). Istotną relacją, jaką przedstawia drzewo jest dość intuicyjny związek pomiędzy populacją i PKB *per capita* a wysokością napływu FDI do kraju. Kraje ludne – a więc o dużym rynku zbytu i potencjale siły roboczej – przyciągają zagraniczne inwestycje i zostały zakwalifikowane jako kraje o „wysokim" napływie FDI. Podobnie ma się rzecz z państwami o wysokim PKB *per capita*, z czego wynika, że kraje rozwinięte, stabilne i ekonomicznie bezpieczne również są atrakcyjne z punktu widzenia inwestorów. Źródeł tej prawidłowości można doszukiwać się utożsamiając wysokie PKB na mieszkańca z wyższą jakością siły roboczej (także wyższą produktywnością), wyższym poziomem wykształcenia oraz lepszym zapleczem technologicznym i infrastrukturalnym. Duża populacja kraju wiąże się zaś z chłonnym rynkiem wewnętrznym i potencjałem gospodarczym.

Otrzymany model potwierdza wpływ liberalizacji handlu zagranicznego na atrakcyjność inwestycyjną. W drzewie J48 pojawiają się zmienne określające łatwość handlu, jak na przykład ilość dokumentów niezbędna przy przywozie i wywozie dóbr. Więcej niż 7 (odpowiednio 9) takich dokumentów wiąże się z kategorią niskich FDI, w przeciwnym razie – średnich. Te zmienne występują jedynie w prawym poddrzewie, gdzie zgrupowano kraje o wartości indeksu informacji kredy-

towej nie mniejszej niż 3 na 6 punktów możliwych. Przedstawiono to rys. 3., gdzie znajduje się przytoczony fragment drzewa decyzyjnego J48.



Źródło: opracowanie własne.

Rys. 3. Fragment drzewa decyzyjnego J48

Kolejnym istotnym czynnikiem jest *Credit Information Index*. W drzewie J48 indeks informacji kredytowej znajduje się w samym korzeniu drzewa, jest zatem zmienną na której przeprowadza się pierwszy test dzielący obserwacje na dwie podgrupy. Do lewej gałęzi trafiają obserwacje, dla których wartość indeksu jest niższa bądź równa 3 (na skali od 0 do 6), do prawej zaś instancje, dla których jest ona wyższa od 3 (Tabela 5.).

Tabela 5. Poziom napływów FDI w zależności od indeksu informacji kredytowej

Wartość indeksu informacji kredytowej	**Poziom napływu bezpośrednicz inwestycji zagranicznych**						
	niski		**średni**		**wysoki**		Ogółem
	Liczebność	Liczebność jako % z wiersza	Liczebność	Liczebność jako % z wiersza	Liczebność	Liczebność jako % z wiersza	Liczebność
(0;3>	245	52%	170	36%	57	12%	472
(3;6>	39	10%	114	30%	226	60%	379

Źródło: opracowanie własne.

Co ważne, jak pokazuje Tabela 5., do lewej gałęzi trafia relatywnie więcej instancji zaklasyfikowanych jako „niskie FDI" zaś do prawej jako „wysokie". Można

wnioskować, że w państwach z wysokimi napływami FDI, indeks ten osiąga wartości powyżej 3, to znaczy powyżej połowy możliwych do zdobycia punktów.

Kolejne gałęzie drzewa decyzyjnego ukazują wpływ na FDI takich grup zmiennych jak proces nabycia i rejestracji nieruchomości, liczba procedur i biurokracja, opodatkowanie czy też egzekucja postanowień kontraktowych.

Podsumowanie

Jak widać z zaprezentowanego badania, drzewo decyzyjne sprawdza się jako metoda analizy dużych zestawów danych. Atutem nie do przecenienia jest szybkość wygenerowania modelu przy jednoczesnej dbałości o wysoką jakość otrzymanych wyników. Ponadto drzewo stanowi ciekawą wizualizację zauważonych relacji, stąd może być wykorzystywane przy rozwiązywaniu problemów biznesowych, kiedy przejrzystość jest równie ważna co dokładność. Data mining oferuje także inne sposoby analizy, na przykład za pomocą metod decyzyjnych, które również warte są uwagi. Wybór jednej najlepszej dla opisu zadanego zagadnienia często zależy od efektów, które chce się uzyskać oraz grupy docelowych odbiorców wyniku przeprowadzonej analizy.

Piśmiennictwo

1. Blonigen B. A., 2005. *A Review of the Empirical Literature on FDI Determinants*; University of Oregon and NBER.
2. Moosa I. A., 2009. *Foreign Direct Investment: Theory, Evidence and Practice*, Palgrave.
3. Mankiw G., 2009. *Principles of Macroeconomics*; South-Western Cengage Learning.
4. Mankiw G., 2009. *Macroeconomics*, Worth Publishers.
5. *OECD Benchmark Definition of Foreign Direct Investment – Third Edition*, OECD, Paris 1996.
6. Witten I. H., Eibe F., 2005. *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann.
7. Raport Doing business, http://www.doingbusiness.org.
8. Strona Banku Światowego, http://www.worldbank.org.