Szkoła Główna Handlowa w Warszawie
Kolegium Analiz Ekonomicznych
Instytut Informatyki i Gospodarki Cyfrowej

mgr Przemysław Pospieszny

Autoreferat rozprawy doktorskiej pt.:

Zastosowanie technik eksploracji danych do estymacji pracochłonności i czasu trwania projektów informatycznych

Praca doktorska napisana pod kierunkiem naukowym
dr. hab. Andrzeja Kobylińskiego, prof. nadzw. SGH

Warszawa 2015

1. Zarys problematyki

W niniejszej rozprawie zostało zaproponowane alternatywne podejście do estymacji pracochłonności i czasu trwania inicjatyw informatycznych w stosunku do dostępnych tradycyjnych metod szacowania, opartych w znacznym stopniu na wiedzy eksperckiej oraz na liniach kodu źródłowego i punktach funkcyjnych. Wykorzystano do tego techniki *data mining* wywodzące się ze statystyki, uczenia maszynowego i sztucznej inteligencji. Umiejętności predykcyjne algorytmów eksploracji danych są powszechnie uznane, co przejawia się w coraz większym zakresie ich aplikacji w ostatnich dwóch dekadach w różnych sektorach gospodarki. Szczególne zastosowanie znajdują one w dziedzinach charakteryzujących się dużą złożonością i niepewnością co do produktu lub rezultatu końcowego. Dlatego też stosowane są do takich zagadnień, jak ocena ryzyka kredytowego, zarządzanie relacjami z klientem czy też wykrywanie nadużyć.

Koncepcja zastosowania technik eksploracji danych do estymacji pracochłonności i czasu trwania projektów informatycznych wywodzi się z problematyki dużego poziomu inicjatyw, będącego według niektórych badań nawet na poziomie 65%[1], niedotrzymujących ustalonego w fazie inicjacji oraz planowania kosztu i czasu ich realizacji. W rezultacie, produkty będące wynikiem realizacji projektów informatycznych, często odbiegają od założonego zakresu funkcjonalnego i charakteryzują się niską jakością, co prowadzi do niezadowolenia klienta końcowego. Dodatkowo, wyższy od założonego budżet i czas trwania potrzebny do wytworzenia produktu mogą powodować niekorzystny bilans zysków biznesowych w stosunku do poniesionych kosztów, prowadząc do przedwczesnego zaniechania realizacji rozpoczętych już projektów.

Przyczyn zaniechania projektów upatruje się zazwyczaj nie na etapie błędów poczynionych podczas budowy produktu końcowego, testów lub wdrożenia, lecz w fazie inicjacji i planowania[2]. Wspomniane założenia projektu raz sprecyzowane na początku inicjatywy, w późniejszych fazach tylko w rzadkich przypadkach ulegają zmianie. Wymagane jest przy tym

[1] Standish Group, *The CHAOS Manifesto 2011*, „The Standish Group International. EUA", 2011; B. Czarnacka-Chrobot, *Analysis of the functional size measurement methods usage by Polish business software systems providers*, „Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)", 2009, t. 5891 LNCS, s. 17–34.
[2] G. Wells, *Why Projects Fail*, „Management Science Journal," 2003.

przeprowadzenie procesu zgłoszenia zmiany, co wiąże się z rewizją rachunku korzyści. W przypadku kosztów przewyższających potencjalne zyski z wdrożenia produktu końcowego przedsięwzięcia, zmiany zgłoszone do harmonogramu lub budżetu inicjatywy mogą nie uzyskać zgody sponsora oraz interesariuszy, a tym samym wpłynąć na zakończenie projektu porażką.

Estymacja założeń projektu, pracochłonności i czasu trwania we wczesnych fazach realizacji przedsięwzięcia jest niezwykle trudnym zadaniem ze względu na niepełną wiedzę, jaka jest dostępna odnośnie do produktu finalnego inicjatywy i zadań związanych z jego realizacją. Błędy popełnione w procesie szacowania mają bezpośredni wpływ na przebieg i zakończenie projektu sukcesem. Jednak częstą praktyką w organizacjach jest estymowanie założeń projektu na podstawie wytycznych kadry zarządczej lub klienta. Dodatkowo, proces szacowania opiera się na metodach tradycyjnych, takich jak estymacja przez analogię, metoda ekspercka lub dekompozycja. W przypadku niedoświadczonych kierowników projektów wynikiem takich działań mogą być zbyt optymistyczne założenia co do pracochłonności i długości trwania inicjatyw, co w późniejszych fazach projektu może prowadzić do porażki, szczególnie gdy koszty przewyższają korzyści z wdrożenia systemu informatycznego.

Organizacje dojrzałe pod względem zarządzania projektami, szczególne te posiadające certyfikację CMMI (ang. *Capability Maturity Model Integration*, model badania i oceny dojrzałości procesowej organizacji)[3], dostrzegają wagę dokładnej estymacji jako warunek zakończenia inicjatywy sukcesem i stosują zaawansowane metody szacowania, opierające się na wymiarowaniu oprogramowania w postaci linii kodu źródłowego (ang. *source line of code*, SLOC) lub punktów funkcyjnych (ang. *function points*, FP). Podejścia te są nieustanie rozwijane od końca lat 70. ubiegłego wieku i zapewniają standaryzację, powtarzalność i ciągłe doskonalenie procesu estymacji projektów przez aktualizację i kalibrację poszczególnych technik. Jednak oba podejścia mają wiele mankamentów, które ograniczają powszechne ich użycie w praktyce. Techniki oparte na SLOC są niedostosowane do współczesnych języków programowania oraz nie uwzględniają pracochłonności innej niż wytwarzanie oprogramowania, takiej jak zbieranie wymagań oraz testy[4]. Natomiast te oparte

[3] A. Kobyliński, *Jakościowe aspekty produkcji oprogramowania*, „Roczniki Kolegium Analiz Ekonomicznych" , Szkoła Główna Handlowa, 1999, z. 7, s. 89–103.
[4] D. Galorath, M. Evans, *Software Sizing, Estimation, and Risk Management*, Auerbach Publications, Boca Raton 2006, s.12.

na punktach funkcyjnych wymagają specyfikacji produktu końcowego, tak więc wyłącznie w ograniczonej postaci mogą być stosowane w początkowej fazie projektu. Dodatkowo opierają się na subiektywnej ocenie estymującego[5]; wielkość systemu, pracochłonność i czas trwania potrzebne na jego wytworzenie mogą być zatem różnie oszacowane w wyniku indywidualnej oceny asesora. Stosowanie zarówno SLOC, jak i FP wymaga dysponowania przeszkolonym personelem, który w dużej mierze wykonuje manualne kalkulacje zarówno do wymiarowania oprogramowania, jak i wyznaczenia parametrów projektu, czego wynikiem mogą być często błędne i nader optymistyczne estymacje względem rzeczywistych wartości.

Środowisko realizacji projektów informatycznych charakteryzuje się dużą zmiennością ze względu na postęp technologiczny[6] obserwowany w całej historii informatyki. Przejawia się on w zwiększającej się złożoności systemów informatycznych, pojawianiu się nowych języków programowania oraz metodyk wytwórczych i zarządzania projektami. Wspomniana zmienność wymaga, aby metody oparte na liniach kodu źródłowego oraz punktach funkcyjnych były poddawane ciągłej adaptacji, co często jest procesem czaso- i pracochłonnym, ze względu na mnogość zależności wpływających na proces estymacji. Efektem są często niedokładne modele predykcyjne, które nie odzwierciedlają praktyk wdrożeniowych oraz kultury organizacyjnej przedsiębiorstw. Dodatkowo zarówno SLOC, jak i FP służą przede wszystkim do wymiarowania oprogramowania i nie mogą być zastosowane do projektów z obszaru zarządzania zmianą (ang. *change management*), które nastawione są nie na wdrożenie konkretnego systemu, lecz wytworzenie metodyki, procesu lub innej zmiany wpływającej na efektywność organizacji, która to zmiana przekłada się na wzrost ekonomicznej wartości danej instytucji.

Z wymienionych powyżej powodów badacze zajmujący się problematyką estymacji projektów informatycznych[7] w ostatnich latach zwrócili uwagę na nową dyscyplinę zajmującą

[5] C. Kemerer, *Reliability of function points measurement: a field experiment*, „Communications of the ACM," 1993, t.36, no 2, s. 85–97.
[6] F.J. Heemstra, *Software cost estimation*, „Information and Software Technology," 1992, t.34, pp. 627–639.
[7] D. Dzega, W. Pietruszkiewicz, *Classification and metaclassification in large scale data mining application for estimation of software projects*, „2010 IEEE 9th International Conference on Cybernetic Intelligent Systems, CIS 2010", 2010; K. Dejaeger et al., *Data mining techniques for software effort estimation: A comparative study*, „IEEE Transactions on Software Engineering", 2012, t. 38, s. 375–397; I.F. de Barcelos Tronto, J.D.S. da Silva, N. Sant'Anna, *Comparison of Artificial Neural Network and Regression Models in Software Effort Estimation*, „Neural Networks, 2007. IJCNN 2007. International Joint Conference on", 2007, s. 771–776.

się eksploracją dużych zbiorów danych: *data mining*[8]. Koncept ten zrodził się na początku lat 90. ubiegłego wieku i miał swoje źródło w popularyzacji hurtowni danych, analityki biznesowej (ang. *business intelligence*) oraz zarządzania wiedzą. Ze względu na łączenie technik wywodzących się z różnych dziedzin nauki, takich jak statystyka, matematyka, czy uczenie maszynowe, algorytmy *data mining* charakteryzują się dużą dokładnością estymacji i znajdują praktyczne zastosowanie w budowaniu modeli zwiększających ekonomiczną wartość organizacji. Przeprowadzone prace badawcze w ostatnim 20-leciu wykazały ich potencjalne możliwości w zakresie estymacji budżetu, harmonogramu oraz jakości produktu końcowego[9] (w odniesieniu do liczby i rodzaju błędów). Dodatkowo, algorytmy *data mining,* zarówno uczenia nadzorowanego, jak i nienadzorowanego, mogą być stosowane jako narzędzie monitorujące postęp w realizacji projektów, przykładowo w analizie *Earned Value Management* (EVM)[10], czy też do predykcji przyszłych kosztów utrzymania systemów [11].

Szczególną uwagę poświęca się szacowaniu pracochłonności i czasu trwania inicjatyw mających na celu wytworzenie lub rozbudowę istniejącego systemu na wstępnym etapie projektu[12]. Stanowi ono największe wyzwanie dla estymujących ze względu na niepełną informację odnośnie do wymagań produktu końcowego, niepewność co do aktywności związanych z jego wytworzeniem oraz dużym prawdopodobieństwem materializacji ryzyk. W tym celu stosowano indywidualne algorytmy predykcyjne, które generowały estymacje pracochłonności i czasu trwania inicjatyw, tym samym wykazując ich przydatność do przedstawiania badanych zjawisk.

Jednak dotychczas techniki eksploracji danych nie znalazły powszechnego zastosowania w praktyce w organizacjach budujących systemy informatyczne jako narzędzie wspierające

[8] Termin *data mining* w literaturze polskojęzycznej bywa tłumaczony jako eksploracja danych. Jednak terminologia ta nie w pełni przedstawia istotę i techniki multidyscyplinarne stojące za *data mining*.

[9] N.K. Nagwani, A. Bhansali, *A data mining model to predict software bug complexity using bug estimation and clustering*, „ITC 2010 - 2010 International Conference on Recent Trends in Information, Telecommunication, and Computing," 2010, s. 13–17.

[10] S.H. Iranmanesh, Z. Mokhtari, *Application of data mining tools to predicate completion time of a project*, „Proceeding of World Academy of Science, Engineering and Technology," 2008, t.32, s. 234–240.

[11] R. Shukla, A.K. Misra, *Estimating software maintenance effort a neural network approach*, „Proceedings of the 2008 1st India Software Engineering Conference, ISEC'08," 2008, s 107–112.

[12] C. Lopez-Martin, C. Isaza, A. Chavoya, *Software development effort prediction of industrial projects applying a general regression neural network*, „Empirical Software Engineering", 2012, t. 17, s. 738–756; J. Villanueva-Balsera et al., *Effort estimation in information systems projects using data mining techniques*, „Proceedings of the 13th WSEAS International Conference on Computers – Held as part of the 13th WSEAS CSCC Multiconference", 2009, s. 652–657.

proces estymacji zasobów niezbędnych do wytworzenia produktów końcowych. Zjawisko to może wynikać z niespójności prowadzonych w tym zakresie prac badawczych. Otrzymane wyniki różniły się w zależności od użytych technik, ich konfiguracji i wykorzystanej historycznej bazy projektów do procesu uczenia się algorytmów. Stosowano przeważnie indywidualne modele, które wdrożone w praktyce mogą generować odmienne rezultaty niż te otrzymane w pracach badawczych, ze względu na specyfikę danych w wybranej organizacji. Dodatkowo większość modeli była budowana na podstawie zbiorów obserwacji o liczebności mniejszej niż 100, pochodzących od jednej wybranej instytucji. Stąd możliwość przeuczenia i zaburzenia rzeczywistej zdolności predykcyjnej algorytmów. Innym aspektem jest jakość danych o projektach, która w wielu organizacjach jest na niskim poziomie. Tymczasem proponowane przez badaczy algorytmy powinny być odporne na brakujące wartości i szumy w danych. Powyższe niespójności w pracach badawczych, brak propozycji zintegrowanego podejścia do szacowania projektów informatycznych oraz niepopularność stosowania tych metod w praktyce przyczyniły się do podjęcia tej tematyki w niniejszej rozprawie doktorskiej.

2. Przedmiot rozprawy i metodyka badawcza

Przedmiotem rozprawy jest wykorzystanie agregacyjnych predykcyjnych technik eksploracji danych do estymacji pracochłonności i czasu trwania projektów informatycznych na ich początkowym etapie, celem opracowania modeli mogących zostać potencjalnie wykorzystanych w praktyce. Do przeprowadzenia badań wybrano bazę danych ISBSG[13], zawierającą historyczne dane o zakończonych inicjatywach informatycznych pochodzących od wielu instytucji publicznych i prywatnych, działających w różnych gałęziach przemysłu i administracji, dotyczących zarówno wytworzeniu nowego oprogramowania, jak i modyfikacji istniejącego. Budowa wspomnianych modeli eksploracji danych odbyła się zgodnie z uznaną metodyką *Cross Industry Standard Process for Data Mining* (CRISP-DM)[14]. Najpierw przeprowadzono analizę danych, celem wyłonienia podzbioru uczącego, i analizę zależności pomiędzy zmiennymi, a także ich wpływu na zmienne zależne pracochłonność i czas trwania. Do budowy modeli wybrano trzy predykcyjne algorytmy eksploracji danych: ogólne modele liniowe (ang. *generalized linear model*, GLM), wielowarstwowe sieci neuronowe (ang.

[13] International Software Benchmarking Standards Group, *ISBSG Repository Data Release 12 - Field Descriptions*, 2013.
[14] C. Pete et al., *CRISP-DM 1.0*, *CRISP-DM Consortium*, 2000.

multilayer perceptron artificial neural network, MLP) oraz drzewa decyzyjne CHAID (ang. *CHi-squared Automatic Interaction Detection decision trees*, CHAID). Wybór tych algorytmów nastąpił na podstawie przeglądu literatury oraz wyników wstępnej analizy modeli, które najlepiej odwzorowywały badane zjawiska oraz charakteryzowały się odpornością na brakujące i zaszumione dane. Techniki te zostały użyte do budowy dwóch modeli, oddzielnie dla zmiennych zależnych pracochłonność i czas trwania. Każdy z modeli składa się z trzech wspomnianych algorytmów, których wyniki zostały uśrednione, zgodnie z najlepszą praktyką, celem otrzymania dokładniejszej estymacji. Ewaluacja modeli miała za zadanie potwierdzić zdolność do predykcji badanych zjawisk, większą dokładność estymacji agregacyjnego modelu niż indywidualnych technik oraz możliwość implementacji modeli w praktyce.

Wyżej przedstawione podejście umożliwia w praktyce estymację projektów informatycznych związanych nie tylko z wytworzeniem lub rozbudową istniejącego systemu, lecz także z wdrożeniem lub zmianą metodyki pracy, procesów i procedur realizowanych w ramach środowiska zarządzania projektami informatycznymi. Agregacyjne techniki pozwalają organizacjom na sprawne wdrożenie modeli w praktyce, co będzie polegać na przygotowaniu danych i dopasowaniu wytworzonych na potrzeby tej rozprawy modeli do charakterystyki inicjatyw w ramach danej instytucji. W rezultacie połączenia zdolności predykcyjnej trzech efektywnych algorytmów, zamiarem autora rozprawy było otrzymanie narzędzia odpornego na słabą jakość danych wejściowych i unikatowość poszczególnych organizacji, pod względem kultury pracy i danych. Modele zostały zbudowane z użyciem narzędzia IBM SPSS Modeler, które jest powszechnie wykorzystywane przez instytucje stosujące techniki eksploracji danych. Rezultatem dodatkowym tej rozprawy jest sformułowanie propozycji metodyki wdrożenia modeli w praktyce.

3. Cele i hipotezy badawcze rozprawy

Celem głównym pracy była budowa agregacyjnych modeli predykcyjnych z użyciem ogólnych modeli liniowych, sieci neuronowych oraz drzew decyzyjnych CHAID do estymacji pracochłonności i czasu trwania projektów informatycznych. Poniżej przedstawiono

poboczne cele badawcze, umożliwiające realizację założonego celu głównego i sformułowanych hipotez.

Cele poznawcze:

- Określenie zależności pomiędzy zmiennymi opisującymi projekty informatyczne oraz ich wpływu na szacowanie pracochłonności i czasu trwania inicjatyw.
- Ocena przydatności ogólnych modeli liniowych, wielowarstwowych sieci neuronowych oraz drzew decyzyjnych CHAID do estymacji pracochłonności i czasu trwania projektów informatycznych.

Cele metodyczne:

- Opracowanie podejścia do budowy agregacyjnych predykcyjnych modeli eksploracji danych estymujących pracochłonność i czas trwania projektów informatycznych z użyciem trzech technik regresyjnych *data mining*: ogólnych modeli liniowych, wielowarstwowych sieci neuronowych oraz drzew decyzyjnych CHAID.
- Zaproponowanie metodyki wdrożenia zbudowanych modeli w praktyce.

Cel aplikacyjny:

- Budowa agregacyjnego modelu estymującego pracochłonność i czas trwania inicjatyw z użyciem wielobranżowej bazy historycznych projektów informatycznych, celem wstępnej kalibracji algorytmów predykcyjnych oraz oceny ich możliwości aplikacji, w rezultacie potencjalnego wdrożenia modelu w ramach procesów zarządzania inicjatywami w różnego typu organizacjach realizujących projekty informatyczne.

W niniejszej rozprawie sformułowano następującą hipotezy badawcze:

1. Predykcyjne techniki eksploracji danych (*data mining*) mogą znajdować zastosowanie w zarządzaniu projektami informatycznymi, wspomagając proces estymacji pracochłonności i czasu trwania inicjatyw na ich inicjalnym etapie oraz potencjalnie przyczyniać się do wzrostu prawdopodobieństwa zakończenia projektu sukcesem. Przez to stanowią one narzędzie konkurencyjne do metod tradycyjnych oraz metod wykorzystujących linie kodu źródłowego lub punkty funkcyjne.
2. Ogólne modele linowe, wielowarstwowe sieci neuronowe oraz drzewa decyzyjne CHAID charakteryzują się dostatecznie dobrą zdolnością predykcyjną

pracochłonności i czasu trwania projektów informatycznych oraz odpornością na braki i szumy w danych, umożliwiając potencjalne ich wdrożenie w praktyce.

3. Agregacyjne predykcyjne modele eksploracji danych zastosowane do estymacji projektów informatycznych na początkowym etapie umożliwią otrzymywanie dokładniejszych szacunków badanych zjawisk niż użyte indywidualnie algorytmy.

4. Zawartość i układ rozprawy

Rozprawa doktorska została podzielona na cztery rozdziały. W rozdziale pierwszym omówiono czynniki wpływające na projekty informatyczne wraz z kryteriami zakończenia ich sukcesem. Stanowią one punkt wyjściowy do zrozumienia problematyki szacowania parametrów przedsięwzięć, ponieważ ich poprawna estymacja wpływa na możliwość wytworzenia produktu finalnego projektu w zgodzie ze zdefiniowaną wstępnie pracochłonnością i długością trwania. Poza tym mają one bezpośredni wpływ na jakość produktu inicjatywy oraz jego odbiór przez klienta lub sponsora. W kolejnych podrozdziałach omówiono problematykę estymacji projektów i miar stosowanych w procesie szacowania. Przedstawione zostały również klasyczne techniki estymacji, takie jak: przez analogię, szacowanie eksperckie czy dekompozycja, jak i bardziej zaawansowane metody parametryczne oparte na liniach kodu źródłowego (COCOMO/ COCOMO II, SLIM, SEER-SEM) oraz na wymiarowaniu oprogramowania z użyciem punktów funkcyjnych (IFPUG, NESMA, COSMIC). Rozdział ten omawia również niedoskonałości stosowanych obecnie technik szacowania projektów informatycznych.

Rozdział drugi poświęcony jest tematyce odkrywania wiedzy w zarządzaniu projektami i technikom eksploracji danych. Dokonano w nim także przeglądu literatury z zakresu zastosowania technik *data mining* do estymacji parametrów projektu. W pierwszej kolejności przedstawione zostały obszary wiedzy wyróżniane w procesie zarządzania projektami oraz opisano informacje zbierane na każdym z etapów w postaci zbiorów danych, zawierających takie charakterystyki przedsięwzięć, jak budżet, pracochłonność, czas trwania, zastosowany język programowania, metodyka realizacji pracy czy też liczba błędów wykrytych podczas fazy testów (jakość). Następnie omówiono proces odkrywania wiedzy w zbiorach danych, którego istotnym krokiem jest eksploracja danych (*data mining*). W poszczególnych

podrozdziałach przedstawiono rodzaje technik *data mining*, metodykę realizacji procesu pozyskiwania wiedzy z danych CRISP-DM oraz trzy algorytmy predykcyjne, które w rozdziałach 3 i 4 zostały wykorzystane do budowy modeli estymujących pracochłonność i czas trwania projektów: ogólne modele liniowe, wielowarstwowe sieci neuronowe oraz drzewa decyzyjne CHAID. W dalszej części rozdziału przeprowadzono przegląd dotychczasowej literatury z zakresu wykorzystania technik eksploracji danych do estymacji parametrów projektów informatycznych, wskazując na stosowane podejścia, techniki, bazy danych użyte do procesu uczenia się oraz metody ewaluacji modeli zbudowanych przy użyciu algorytmów *data mining*. Istotnym aspektem tej części pracy są rozważania na temat ograniczeń dotychczasowych badań z obszaru zastosowań modeli predykcyjnych.

Rozdział trzeci rozpoczyna część empiryczną rozprawy, gdzie do przeprowadzenia badań wykorzystano metodykę CRISP-DM. Do budowy modeli predykcyjnych pracochłonności i czasu trwania zdecydowano się zastosować wielobranżową bazę projektów informatycznych ISBSG, zawierającą ponad 6000 projektów, tak aby możliwie jak najlepiej odzwierciedlić różnorodność realizowanych projektów oraz organizacji je wykonujących. Pierwszym krokiem mającym na celu przygotowanie zbioru danych wejściowych do modelowania predykcyjnego pracochłonności i czasu trwania było zrozumienie i analiza danych względem występujących wartości brakujących, przeprowadzenie transformacji danych oraz analizy współzależności Pearsona oraz regresji krokowej. Następnie przedstawiono, oddzielnie dla pracochłonności i czasu trwania inicjatyw, proces budowy oraz konfiguracji modeli agregacyjnych, składających się z trzech technik predykcyjnych: ogólnych modeli linowych, sieci neuronowych oraz drzew decyzyjnych CHAID, wraz z uzasadnieniem ich wyboru.

Rozdział czwarty stanowi ewaluację indywidualnych modeli oraz agregacyjnych (uśrednionych), pod względem dokładności estymacji oraz możliwości wykorzystania ich w praktyce. Do oceny wybrano tradycyjne kryteria błędu prognozy, takie jak: średni błąd, średni absolutny błąd, średni błąd kwadratowy i pierwiastek błędu średniokwadratowego, oraz powszechnie stosowane do ewaluacji modeli szacowania pracochłonność i czas trwania projektów informatycznych: moduł błędu względnego (ang. *mean relative error*, MRE), średni moduł błędu względnego (ang. *mean magnitude of relative error*, MMRE) i stosunek predykcji do wartości rzeczywistych (PRED). W pierwszej kolejności ocenie podlegał każdy z trzech modeli zbudowanych oddzielnie do szacowania pracochłonności i czasu trwania z użyciem ogólnych modeli liniowych, sieci neuronowych oraz drzew decyzyjnych CHAID.

Następnie ewaluacji poddano dwa modele agregacyjne każdy składający się z trzech wspomnianych algorytmów estymujących badane zjawiska. Wyniki zintegrowanego podejścia zostały porównane z zastosowanymi indywidualnymi modelami w celu określenia wyższości przejawiającej się w dokładności wartości szacowanych. Ostatnia część rozdziału czwartego przedstawia propozycję metodyki wdrożenia agregacyjnego modelu w praktyce.

Rozprawę podsumowują wnioski z przeprowadzonych badań oraz odniesienie do założonych celów i hipotez badawczych. Przedstawione są również ograniczenia pracy, a także wskazane potencjalne kierunki przyszłych badań.

W ostatniej części pracy została zwarta literatura źródłowa oraz załączniki prezentujące szczegółowe wyniki otrzymanych modeli predykcyjnych.

5. Wyniki i wnioski z przeprowadzonych badań

Głównym celem pracy była budowa dwóch modeli predykcyjnych (zagregowanych), oddzielnie dla pracochłonności i czasu trwania projektów, z użyciem trzech technik *data mining*, przez uśrednianie wyników estymacji algorytmów eksploracji danych: ogólnych model liniowych, wielowarstwowych sieci neuronowych i drzew decyzyjnych. Do procesu uczenia i walidacji modeli wykorzystano wielobranżową bazę historycznych projektów informatycznych ISBSG, charakteryzującą się dużym wolumenem dobrych jakościowo danych. Stosowana jest ona przez organizacje członkowskie ISBSG do wsparcia procesu wymiarowania nowych inicjatyw. Powyższe podejście umożliwiło budowę modeli, które cechują się bardzo dużą dokładnością szacowania pracochłonności i czasu trwania projektów. Uzyskane miary błędów, przedstawione w tabeli 1 i 2, zarówno dla algorytmów indywidualnych, jak i dla modeli agregacyjnych były na niskim poziomie, spełniając przy tym kryterium Conte'a[15] odnoszące się do średniego modułu błędu względnego, który był na poziomie mniejszym niż 0,25. Niemniej zastosowane podejście agregacyjne oparte na uśrednianiu wyników estymacji poszczególnych technik do szacowania pracochłonności i

[15] S.D. Conte, H.E. Dunsmore, and V.Y. Shen, *Software engineering metrics and models*, Benjamin/Cummings Pub. Co. 1986.

czasu trwania inicjatyw generowało dokładniejsze predykcje dla badanych zjawisk niż indywidualne algorytmy.

Tabela 1 Porównanie modeli indywidualnych oraz agregacyjnego dla pracochłonności

	Ogólny model linowy		Wielowarstwowa sieć neuronowa		Drzewo decyzyjne CHAID		Model agregacyjny	
	Uczenie	**Test**	**Uczenie**	**Test**	**Uczenie**	**Test**	**Uczenie**	**Test**
Błąd minimalny	-1,523	-1,170	-1,507	-1,410	-1,528	-1,334	-1,389	-1,227
Błąd maksymalny	1,172	1,073	1,334	1,192	1,362	1,200	1,230	1,155
ME	0,000	-0,012	0,008	0,002	0,000	-0,008	-0,004	-0,011
MAE	0,288	0,310	0,308	0,331	0,287	0,313	0,288	0,310
MSE	0,139	0,162	0,159	0,175	0,140	0,169	0,139	0,160
RMSE	0,373	0,402	0,398	0,418	0,374	0,412	0,373	0,400
MMRE	0,203	0,053	0,226	0,113	0,225	0,050	0,187	0,040
PRED(0,25)	0,599	0,604	0,571	0,545	0,612	0,607	0,618	0,597
PRED(0,3)	0,680	0,662	0,657	0,623	0,680	0,662	0,685	0,662
Odchylenie standardowe	0,373	0,403	0,398	0,419	0,374	0,412	0,368	0,397
Korelacja liniowa	0,768	0,713	0,729	0,687	0,767	0,698	0,775	0,722

Źródło: Opracowanie własne

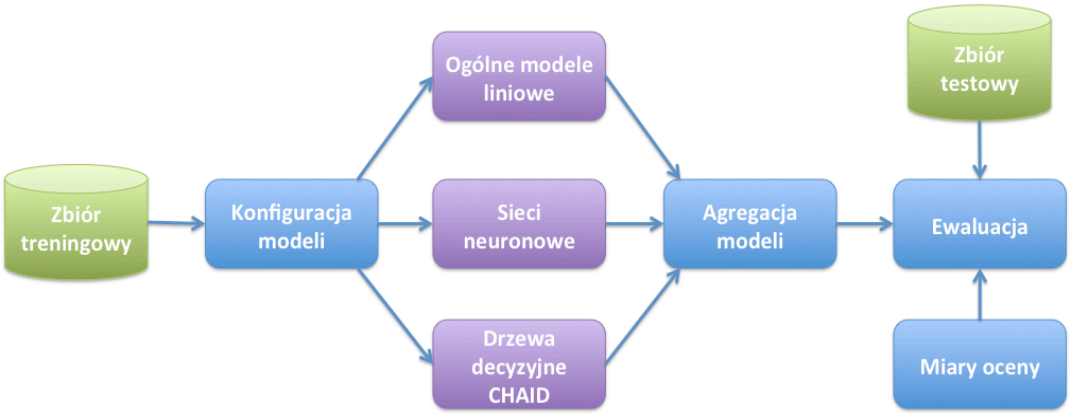
Tabela 1 przedstawia wskaźniki oceny poszczególnych technik oraz modelu agregacyjnego do predykcji roboczomiesięcy niezbędnych do wytworzenia produktu końcowego projektu. Zgodnie z danymi w niej zawartymi, agregacyjny model generuje błędy prognoz na poziomie podobnym do najlepszego z algorytmów, którym był ogólny model liniowy. Jednak analizując miary dopasowania do danych i zdolności estymacji projektów, stosując model agregacyjny osiągnięto najniższy błąd prognozy (MMRE), wynoszący dla zbioru treningowego 0,187 oraz testowego 0,04. Jest to bardzo dobry poziom tego wskaźnika, świadczący o wysokiej jakości modelu, jego zdolności do przewidywania pracochłonności i niskim poziomie generowanych przez niego błędów prognoz. Dodatkowo wskaźnik PRED(0,25) dla modelu agregacyjnego również był na najlepszym poziomie w odniesieniu do modeli indywidualnych i wynosił w przybliżeniu 60% dla obu użytych zbiorów danych (treningowego i testowego).

Tabela 1 Porównanie modeli indywidualnych oraz agregacyjnego dla długości trwania

	Ogólny model linowy		Wielowarstwowa sieć neuronowa		Drzewo decyzyjne CHAID		Model agregacyjny	
	Uczenie	Test	Uczenie	Test	Uczenie	Test	Uczenie	Test
Błąd minimalny	-1,511	-1,085	-1,393	-1,048	-1,452	-0,964	-1,411	-1,032
Błąd maksymalny	0,845	0,702	0,904	0,660	0,971	1,082	0,907	0,804
ME	0,000	0,003	0,000	0,009	0,000	0,012	0,000	0,008
MAE	0,206	0,217	0,188	0,198	0,193	0,212	0,186	0,201
MSE	0,075	0,072	0,065	0,063	0,068	0,074	0,064	0,064
RMSE	0,274	0,268	0,255	0,250	0,261	0,273	0,253	0,252
MMRE	0,228	0,263	0,213	0,259	0,217	0,251	0,205	0,245
PRED(0,25)	0,611	0,558	0,623	0,588	0,654	0,568	0,659	0,591
PRED(0,3)	0,700	0,646	0,706	0,653	0,732	0,653	0,750	0,675
Odchylenie standardowe	0,275	0,268	0,255	0,250	0,261	0,273	0,253	0,253
Korelacja liniowa	0,660	0,658	0,715	0,712	0,700	0,648	0,722	0,706

Źródło: Opracowanie własne

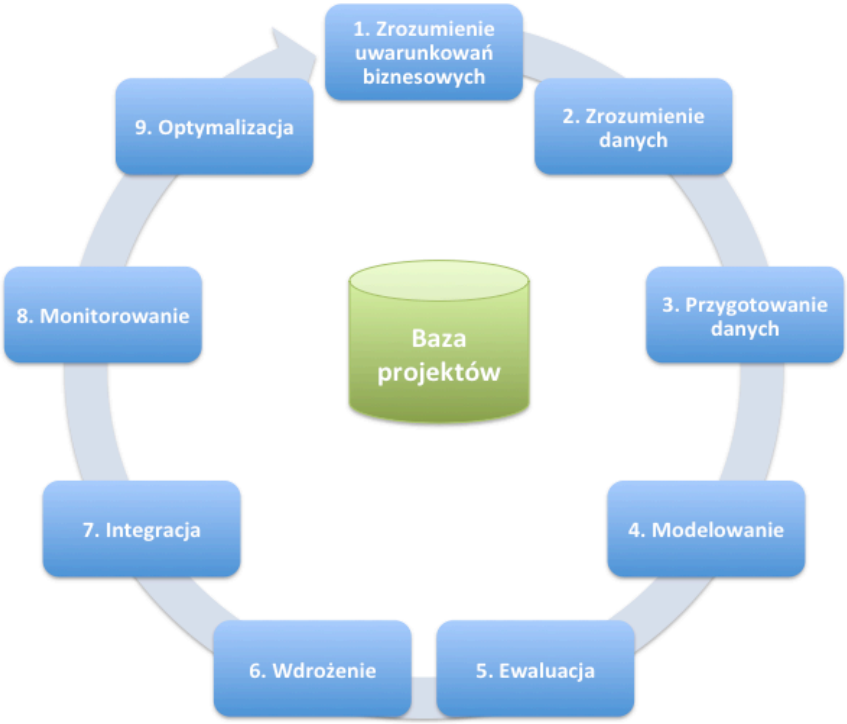
Podobnie jak w przypadku pracochłonności, model agregacyjny do predykcji czasu trwania projektów (tabela 2) generuje lepsze prognozy niż indywidualne techniki. Miary błędu tego modelu ME, MAE i RMSE są zbliżone co do wartości do wyników sieci neuronowej, która jest najbardziej efektywnym algorytmem generującym nieznacznie lepsze szacunki niż pozostałe dwie techniki. Średni moduł błędu względnego (MMRE) dla zintegrowanego modelu w zbiorze treningowym jest na poziomie 0,205 oraz testowym 0,245. Wartości te oznaczają, że agregacyjne techniki mogą generować w przybliżeniu 20-24% błędnych szacunków. Wysokość MMRE na wspomnianym poziomie świadczy o bardzo dobrej zdolności modelu do generowania dokładnych predykcji miesięcy niezbędnych do przeprowadzenia projektu. Wartości PRED(0,25) wynoszą odpowiednio dla zbioru uczącego 66% oraz treningowego 60%.



Rysunek 1 Proces budowy i ewaluacji modeli predykcyjnych

Źródło: Opracowanie własne

Przez realizację wspomnianego celu głównego niniejszej rozprawy osiągnięto cele poboczne, przede wszystkim cel aplikacyjny. Jego istotą było opracowanie modeli *data mining* do estymacji parametrów projektów, umożliwiających dokładne ich szacowanie niezależnie od rodzaju inicjatywy informatycznej oraz organizacji, w której model byłby wdrożony. Cel ten zrealizowano w wyniku zastosowania metody agregacji trzech predykcyjnych algorytmów eksploracji danych wywodzących się z technik opartych na regresji i uczeniu maszynowym: ogólnych modeli liniowych, wielowarstwowych sieci neuronowych i drzew decyzyjnych CHAID (rysunek 1). Zgodnie z uzyskanymi wynikami (tabela 1 i 2) agregacja algorytmów poprzez uśrednianie uzyskiwanych przez nie wyników zwiększyła dokładność estymacji badanych zjawisk i potencjalnie uodporniła modele na szumy w danych, a także wartości nietypowe i odstające. Dodatkowo zniwelowała możliwość wystąpienia efektu nadmiernego dopasowania poszczególnego algorytmu do danych, co na wybranych zbiorach mogłoby generować niepoprawne predykcje. Natomiast zastosowanie bazy ISBSG do procesu uczenia i walidacji modeli, która dostarcza informacji o projektach pochodzących z różnych rodzajów organizacji i typów inicjatyw, umożliwiło ich wstępną kalibrację i potencjalne wdrożenie w dowolnej instytucji realizującej projekty informatyczne. Stanowi to niezwykle istotną właściwość, ponieważ każda organizacja ma odmienną kulturę pracy, metodykę zarządzania projektami oraz poziom kompetencji pracowników. Tym samym zarówno pracochłonność, jak i czas trwania związany z przeprowadzaniem inicjatyw może znacząco różnić się pomiędzy organizacjami przy realizacji tego samego typu projektu informatycznego.



Rysunek 2 Propozycja procesu wdrożenia zaproponowanych modeli predykcyjnych w organizacji

Źródło: Opracowanie własne

W ramach celów metodycznych przedstawiono opracowaną metodykę wdrożenia uzyskanych modeli w organizacjach zainteresowanych ich zastosowaniem do procesu wsparcia decyzyjnego estymacji inicjatyw informatycznych (rysunek 2). Proponowane podejście stanowi rozszerzenie metodyki CRISP-DM o dodatkowe kroki powdrożeniowe, umożlwiające odpowiednie wykorzystanie modeli przez ich integrację z istniejącymi procesami i narzędziami wykorzystywanymi w danej organizacji oraz monitoring i aktualizację. Wymaga ono dostępności dobrej jakościowo bazy historycznych projektów informatycznych charakterystycznych dla danej instytucji, aby dopasować modele do swoistości realizowanych przez nią inicjatyw. Wspomniany zbiór danych powinien być możliwie jak najobszerniejszy, zawierać przynajmniej informacje o 100 zakończonych inicjatywach oraz obejmować zmienne określające: wytworzony produkt finalny (wielkość, architektura, język programowania), zaangażowany zespół projektowy (wielkość, role, kompetencje), zastosowaną metodykę wytwarzania oraz pracochłonność i czas trwania inicjatywy. Umiejętność predykcji parametrów projektu przez modele jest zależna od jakości posiadanych przez organizację danych. Dlatego też zastosowanie zaproponowanej w pracy metody powinno być ograniczone do instytucji mających biuro zarządzania projektami

(PMO) oraz realizujących prace wytwórcze na wysokim poziomie dojrzałości, np. zgodnie z modelem CMMI, co zapewni odpowiedni standard zbierania i aktualizacji bazy inicjatyw. PMO może również odpowiadać za utrzymanie wdrożonych modeli poprzez zapewnienie ich integracji z wykorzystywanymi narzędziami do zarządzania portfelem projektów (ang. *enterprise project management*, EPM), tak aby dostarczać informacji o szacowanych wielkościach inicjatywy przez graficzny interfejs użytkownika. Dodatkowo do jego zadań można zaliczyć monitorowanie dokładności predykcji modeli oraz ich aktualizację względem nowych danych (zakończonych inicjatyw). Umożliwi to zapewnienie poprawności generowanych przez algorytmy estymacji pracochłonności i czasu trwania inicjatyw przede wszystkim w ich wczesnych fazach realizacji (inicjacja lub planowanie).

Wynikiem pobocznym budowy modeli była realizacja celów poznawczych rozprawy, czyli określenie zależności pomiędzy zmiennymi, ich wpływu na predykcję pracochłonności i czasu trwania oraz ocena przydatności wybranych trzech algorytmów do estymacji badanych zjawisk. Baza projektów ISBSG została poddana procesowi przygotowania danych, w którym to dokonano analizy korelacji Spearmana i Pearsona oraz regresji krokowej. Największy wpływ na pracochłonność inicjatyw miała wielkość wytwarzanego produktu końcowego. Pozostałe zmienne wejściowe w mniejszym stopniu wyjaśniały badane zjawisko, jednak na istotnym poziomie. Natomiast w przypadku czasu trwania projektów interakcja pomiędzy zmienną zależną i predyktorami była bardziej równomiernie rozłożona, ze wskazaniem na wielkość systemu, zastosowaną metodykę, typ platformy sprzętowej systemu oraz wymagany poziom dostosowania systemu do wymagań biznesowych. W odniesieniu do przydatności zastosowanych algorytmów do estymacji pracochłonności i czasu trwania inicjatyw zarówno ogólny model liniowy, jak i wielowarstwowe sieci neuronowe oraz drzewa decyzyjne CHAID umożliwiają budowę modeli do predykcji założeń projektów, dostarczających precyzyjnych szacunków przy niewielkim błędzie. Dodatkowo wszystkie wspomniane modele spełniają kryterium MMRE Conte'a, przy nieznacznym odchyleniu od oczekiwanej wartości PRED, czego przyczyną mogło być zastosowanie dużego wolumenu (1494) zróżnicowanych obserwacji do procesu uczenia i walidacji algorytmów.

Realizacja celu głównego oraz pobocznych (aplikacyjne, metodyczne i poznawcze) umożliwiła zweryfikowanie i potwierdzenie zdefiniowanych w niniejszej pracy trzech hipotez badawczych. Zgodnie z wynikami przedstawionymi w tabeli 1 i 2 predykcyjne techniki eksploracji danych znajdują zastosowanie w zarządzaniu projektami do estymacji

pracochłonności i czasu trwania inicjatyw. Z pewnością mogą być wykorzystane jako narzędzie wsparcia decyzyjnego i stanowić uzupełnienie względem tradycyjnych oraz parametrycznych technik estymacji. Umożliwiają dokładniejsze formułowanie założeń projektu, zwiększając prawdopodobieństwo zakończenia inicjatyw sukcesem. Ogólny model liniowy, wielowarstwowe sieci neuronowe oraz drzewa decyzyjne CHAID charakteryzują się bardzo dobrą zdolnością predykcyjną względem badanych zjawisk, a ich agregacja zapewnia otrzymywanie dokładniejszych szacunków oraz przeciwdziała możliwości wystąpienia nadmiernego dopasowania danego algorytmu do danych.

Dotychczasowe opracowania z zakresu wykorzystania technik *data mining* do estymacji projektów informatycznych, przedstawione w przeglądzie literatury, były w przeważającej mierze poświęcone zagadnieniu efektywności poszczególnych algorytmów, stosując przy tym bazy danych często pochodzące sprzed 20-30 lat i zawierające niewielką liczbę obserwacji. Dodatkowo prezentowane w nich wyniki były często niespójne ze względu na zastosowanie różnych podejść do przygotowania danych oraz procesu ich budowy. Dlatego też dotychczas nie odnotowano ich praktycznego użycia w organizacjach realizujących projekty informatyczne. W rezultacie zaproponowane w niniejszej rozprawie podejście, oparte na zespoleniu trzech efektywnych algorytmów eksploracji danych, ich wstępnej kalibracji na podstawie wielobranżowej bazy historycznych projektów pochodzących z ostatniej dekady oraz metodyka wdrożenia i utrzymania, umożliwią ich łatwiejszą implementację w praktyce. Zbudowane modele do predykcji pracochłonności i czasu trwania inicjatyw stanowią alternatywną lub uzupełniającą metodę estymacji parametrów projektu, względem tradycyjnych lub opartych na punktach funkcyjnych i liniach kodu źródłowego. W odróżnieniu od dostępnych podejść, stanowią one automatyczne, niewymagające znacznej pracochłonności, narzędzie wsparcia decyzyjnego, dostarczające dokładnych szacunków parametrów inicjatyw. Dokonanie estymacji sprowadza się do uwzględnienia w modelach charakterystyk nowej, uprzednio nieznanej inicjatywy. Dodatkowym atutem modeli jest łatwość ich implementacji oraz późniejszego utrzymania. Wymagają one jedynie okresowej aktualizacji o nowo zakończone inicjatywy, tak aby dopasować modele do zmieniającej się w organizacji specyfiki prowadzenia projektów informatycznych oraz zapewnić dokładność estymacji pracochłonności i czasu trwania inicjatyw w ich wczesnych fazach życia.

Podsumowując niniejszą rozprawę, należy również wspomnieć o jej ograniczeniach, które mogą zostać wyeliminowane w efekcie przyszłych badań. Zarówno do budowy, jak i do

ewaluacji modeli zastosowano jedną bazę ISBSG, dzieląc ją przy tym na zbiór treningowy i testowy. Przyczyną tego był brak innego wiarygodnego dostępnego zbioru danych, który mógłby zostać użyty do weryfikacji dokładności szacunków uzyskanych wybranymi algorytmami. Preferowanym rozwiązaniem byłaby implementacja opracowanych modeli do estymacji pracochłonności i czasu trwania projektów w jednej wybranej lub wielu organizacjach, w których ich umiejętność przedstawiania badanych zjawisk mogłaby zostać potwierdzona w praktyce. W ten sposób zostałaby zweryfikowana również zaproponowana metodyka wdrożeniowo-utrzymaniowa. Innym ograniczeniem tej rozprawy, które również można wykorzystać w kolejnych badaniach, jest sposób wyznaczania zmiennej dyskretnej przedstawiającą rozmiar wytwarzanego produktu, która w największym stopniu ze zbioru danych oddziałuje na badane zjawiska. Jej wartości w bazie ISBSG opierają się na przedziałach obliczonych metodami FSM. Do wyznaczenia tej wielkości w równym stopniu mogą zostać użyte tradycyjne techniki estymacji, takie jak ekspercka, czy przez analogię. Jednak metoda punktów funkcyjnych powszechnie uważana jest za najbardziej dokładną technikę określania rozmiaru oprogramowania. Dlatego też w przyszłych badaniach mogłoby zostać wypracowane szczegółowe podejście do estymacji parametrów projektu informatycznego wykorzystujące punkty funkcyjne do wyznaczenia rozmiaru systemu oraz zaproponowane w niniejszej rozprawie techniki eksploracji danych do szacowania pracochłonności i czasu trwania niezbędnego do realizacji przedsięwzięcia. Dodatkowo porównaniu mógłby zostać poddany wpływ rozmiaru oszacowanego różnymi metodami FSM na jakość uzyskiwanych predykcji technikami *data mining* odnoszących się do pracochłonności i czasu trwania projektów, celem wyłonienia metody FSM, która w połączeniu z technikami eksploracji danych dostarcza najdokładniejszych estymacji wspomnianych zjawisk.

Bibliografia – wybrane pozycje

Albrecht A.J., Gaffney J.E.. J., *Software Function, Source Lines of Code, and Development Effort Prediction: A Software Science Validation*, „IEEE Transactions on Software Engineering," 1983, t.SE-9.

Azzeh M., Cowling P.I., Neagu D., *Software stage-effort estimation based on association rule mining and Fuzzy set theory*, „Proceedings - 10th IEEE International Conference on Computer and Information Technology, CIT-2010, 7th IEEE International Conference on Embedded Software and Systems, ICESS-2010, ScalCom-2010," 2010, s. 249–256.

Balsera J.V., Montequin V.R., Fernandez F.O., González-Fanjul C.A., *Data Mining Applied to the Improvement of Project Management*, „InTech," 2012.

De Barcelos Tronto I.F., da Silva J.D.S., Sant'Anna N., *Comparison of Artificial Neural Network and Regression Models in Software Effort Estimation*, „Neural Networks, 2007. IJCNN 2007. International Joint Conference on," 2007, s. 771–776.

Boehm B.W., *Software Engineering Economics*, „Prentice Hall.," 1981, t. 10, s. 4 –21.

Cios K., Pedrycz W., Swiniarski R., Kurgan L., *Data Mining A Knowledge Discovery Apporach*, Springer Science, New York, New York, USA 2007.

Clarke B., Fokoue E., Zhang H.H., *Principles and Theory for Data Mining and Machine Learning*, Springer Science, New York, New York, USA 2009.

Conte S.D., Dunsmore H.E., Shen V.Y., *Software engineering metrics and models*, Benjamin/Cummings Pub. Co. 1986.

Czarnacka-Chrobot B., *Analysis of the functional size measurement methods usage by Polish business software systems providers*, „Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)," 2009, t.5891 LNCS, s. 17–34.

Czarnacka-Chrobot B., *Effectiveness of Business Software Systems Development and Enhancement Projects versus Work Effort Estimation Methods*, „International Journal of Social, Management, Economics and Business Engineering," 2013, t.7, nr 9, s. 1329–1336.

Dejaeger K., Verbeke W., Martens D., Baesens B., *Data mining techniques for software effort estimation: A comparative study*, „IEEE Transactions on Software Engineering," 2012, t.38, s. 375–397.

Dzega D., Pietruszkiewicz W., *Classification and metaclassification in large scale data mining application for estimation of software projects*, „2010 IEEE 9th International Conference on Cybernetic Intelligent Systems, CIS 2010," 2010.

Fayyad U., Piatetsky-Shapiro G., Smyth P., *From data mining to knowledge discovery in databases*, „AI magazine," 1996, s. 37–54.

Flasiński M., *Zarządzanie projektami informatycznymi*, Wydawnictwo Naukowe PWN 2013.

Galorath D., Evans M., *Software Sizing, Estimation, and Risk Management*, Auerbach Publications 2006.

Gasik S., *A model of project knowledge management*, „Project Management Journal," 2011, t.42, nr 3, s. 23–44.

Gatnar E., *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa 2008.

Giudici P., Figini S., *Applied Data Mining for Business and Industry*, John Wiley & Sons 2009.

Han J., Kamber M., Pei J., *Data Mining: Concepts and Techniques*, Morgan Kaufmann 2006.

Hand D.J., Mannila H., Smyth P., *Eksploracja danych*, Wydawnictwa Naukowo-Techniczne 2005.

Hill P., *Practical Software Project Estimation: A Toolkit for Estimating Software Development Effort & Duration*, McGraw Hill Professional 2010.

Iranmanesh S.H., Mokhtari Z., *Application of data mining tools to predicate completion time of a project*, „Proceeding of world academy of science, engineering and technology," 2008, t.32, s. 234–240.

Jaszkiewicz A., *Inżyneria oprogramowania*, Helion, Gliwice 1997.

Jorgensen M., Shepperd M., *A Systematic Review of Software Development Cost Estimation Studies*, „IEEE Transactions on Software Engineering," 2007, t.33, nr 1, s. 33–53.

Kemerer C., *Reliability of function points measurement: a field experiment*, „Communications of the ACM," 1993, t.36, nr 2, s. 85–97.
Kisielnicki J., *Zarządzanie wiedzą we współczesnych organizacjach*, Wyższa Szkoła Handlu i Prawa im. Ryszarda Łazarskiego 2003.
Kisielnicki J., Zarządzanie projektami, Wydawnictwo JAK, Warszawa 2011.
Kobyliński A., *Miary procesu i produktu programowego*, „Współczesne kierunki rozwoju informatyki (PTI)," 2004, s. 105–109.
Kobyliński A., Pospieszny P., *Zastosowanie technik eksploracji danych do estymacji pracochłonności projektów informatycznych*, „Studia i Materiały Polskiego Stowarzyszenia Zarządzania Wiedzą," Bydgoszcz 2015, s. 67–82.
Laird L.M., Brennan M.C., *Software Measurement and Estimation: A Practical Approach*, John Wiley & Sons 2006.
Larose D.T., *Data Mining Methods and Models*, John Wiley & Sons 2007.
Linoff G.S., Berry M.J.A., *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, John Wiley & Sons 2011.
Lopez-Martin C., Isaza C., Chavoya A., *Software development effort prediction of industrial projects applying a general regression neural network*, „Empirical Software Engineering," 2012, t.17, s. 738–756.
Marchewka J., *Information Technology Project Managment - Providing Measurable Organizational Value*, *Management*, Wiley 2003.
McConnell S., *Software Estimation: Demystifying the Black Art: Demystifying the Black Art*, Microsoft Press 2009.
Kearns M, Vailiant L., *Cryptographic Limitations on Learning Boolean Formulae and Finite Automata*, „Symposium on Theory of computing (ACM)," 1989, s. 433–444.
Mittas N., Angelis L., *Ranking and clustering software cost estimation models through a multiple comparisons algorithm*, „IEEE Transactions on Software Engineering," 2013, t.39, nr 4, s. 537–551.
Nagwani N.K., Bhansali A., *A data mining model to predict software bug complexity using bug estimation and clustering*, „ITC 2010 - 2010 International Conference on Recent Trends in Information, Telecommunication, and Computing," 2010, s. 13–17.
Neimat T. Al, *Why IT projects fail*, „The project perfect white paper collection," 2005, s. 1–8.
Nonaka I., *A Dynamic Theory of Organizational Knowledge Creation*, „Organization Science," 1994, t.5, nr 1, s. 14–37.
Pai D.R., McFall K.S., Subramanian G.H., *Software effort estimation using a neural network ensemble*, „Journal of Computer Information Systems," 2013, t.53, s. 49–58.
Paliwal M., Kumar U., *Neural networks and statistical techniques: A review of applications*, „Expert Systems with Applications", 2009, t.36, s. 2–17.
Peeters P., Asperen J. van, Jacobs M., Vonk H., Others A., *The application of Function Point Analysis (FPA) in the early phases of the application life cycle A Practical Manual: Theory and case study*, NESMA 2005.
Perechuda K., *Zarządzanie wiedzą w przedsiębiorstwie*, Wydawnictwo Naukowe PWN 2005.
Piatetsky-Shapiro G., Frawley W.J., *Knowledge Discovery in Databases*, *Library Trends*, 1991, t. 48.
Pospieszny P., Czarnacka-Chrobot B., Kobyliński A., *Application of Function Points and Data Mining Techniques for Software Estimation - A Combined Approach*, „25th International Workshop on Software Measurement and 10th International Conference on Software Process and Product Measurement", Springer 2015, s. 96–113.
Resolution Project, *CHAOS Summary 2009*, „Chaos," 2009, s. 1–4.
Ruan D., Chen G., Kerre E.E., *Intelligent data mining: techniques and applications*, Springer Science & Business Media 2005, 5. wyd.
Ruchika Malhotra A.J., *Software Effort Prediction using Statistical and Machine Learning Methods*, „International Journal of Advanced Computer Science and Applications (IJACSA)," 2011, t. 2.
Schapire R.E., *The strength of weak learnability*, „Machine Learning," 1990, t.5, nr 2, s. 197–227.
Schwalbe K., *Information Technology Project Management*, *Technology*, Course Technology, Boston 2014, t. 1.

Selby R.W., *Software engineering: The legacy of Barry W. Boehm*, „Proceedings - International Conference on Software Engineering," 2007, s. 37–38.

Shukla R., Misra A.K., *Estimating software maintenance effort a neural network approach*, „Proceedings of the 2008 1st India Software Engineering Conference, ISEC'08," 2008, s. 107–112.

Sobczyk M., *Statystyka*, PWN, Warszawa 2000.

Spalek S.J., *Critical Success Factors in Project Management -- To Fail or Not To Fail, That is the Question!*, „PMI Global Congress Proceedings," 2005, s. 1–7.

Standish Group, *The CHAOS Manifesto 2011*, „The Standish Group International. EUA," 2011, s. 25.

Szupiluk R., *Dekompozycje wielowymiarowe w agregacji predykcyjnych modeli Data Mining*, Oficyna wydawnicza SGH, Warszawa 2013.

Taylor J., *Wstęp do Analizy Błędu Pomiarowego*, Wydawnictwo Naukowe PWN, Warszawa 1995.

Trendowicz A., *Software Project Effort Estimation: Foundations and Best Practice Guidelines for Success*, 2014.

Trocki M., Grucza B., Ogonek K., *Zarządzanie projektami*, Polskie Wydawnictwo Ekonomiczne 2009.

Villanueva-Balsera J., Ortega-Fernandez F., Rodríguez-Montequín V., Concepción-Suárez R., *Effort estimation in information systems projects using data mining techniques*, „Proceedings of the 13th WSEAS International Conference on Computers - Held as part of the 13th WSEAS CSCC Multiconference," 2009, s. 652–657.

Weiß C., Premraj R., Zimmermann T., Zeller A., *How long will it take to fix this bug?*, „Proceedings - ICSE 2007 Workshops: Fourth International Workshop on Mining Software Repositories, MSR 2007," 2007.

Wen J., Li S., Lin Z., Hu Y., Huang C., *Systematic literature review of machine learning based software development effort estimation models*, „Information and Software Technology," 2012, t.54, s. 41–59.

Xu L., Krzyzak A., Suen C.Y., *Methods of combining multiple classifiers and their applications to handwriting recognition*, „IEEE Transactions on Systems, Man, and Cybernetics," 1992, t.22, nr 3, s. 418–435.

Zhou Z.-H., *Ensemble Methods: Foundations and Algorithms*, CRC Press, Boca Raton 2012.