

Marcin MATUSZNY¹, Izabela KUTSCHENREITER-PRASZKIEWICZ²

INDUKCJA DRZEW DECYZYJNYCH DLA BUDOWY BAZ WIEDZY PROCESÓW PRODUKCYJNYCH

Streszczenie: W artykule przedstawiono proces pozyskiwania wiedzy w oparciu o indukcję drzew decyzyjnych, w sposób graficzny zilustrowano różnice pomiędzy pozyskiwaniem wiedzy w sposób tradycyjny, od eksperta, a także procesem pozyskiwania wiedzy wspomaganym metodami uczenia maszynowego. Omówiono i wyszczególniono metody pozyskiwania wiedzy. W części praktycznej przedstawiono wykorzystanie oprogramowania DeTreeX 4.0 dedykowanego do indukcji drzew decyzyjnych wchodzącego w skład pakietu sztucznej inteligencji Sphinx 4.0.

Słowa kluczowe: drzewa decyzyjne, procesy produkcyjne, bazy wiedzy, pozyskiwanie wiedzy, inżynieria produkcji

INDUCTION OF DECISION TREES FOR BUILDING KNOWLEDGE BASES OF PRODUCTION PROCESSES

Summary: The article presents the process of acquiring knowledge based on the induction of decision trees, graphically illustrating the differences between acquiring knowledge in a traditional way, from an expert, and the process of acquiring knowledge supported by machine learning methods. The methods of acquiring knowledge are discussed and specified. The practical part presents the use of DeTreeX 4.0 software dedicated to the induction of decision trees, which is part of the Sphinx 4.0 artificial intelligence package.

Keywords: decision trees, production processes, knowledge bases, knowledge acquisition, production engineering

1. Pozyskiwanie wiedzy z procesów produkcyjnych dla budowy baz wiedzy

Utworzenie bazy wiedzy systemu eksperckiego wspomagającego projektowanie procesów technologicznych obróbki w wielu przypadkach wymaga opracowania metod pozyskiwania wiedzy technologicznej. Proces pozyskiwania wiedzy technologicznej ma przede wszystkim na celu pozyskanie wiedzy oraz doświadczenia

¹ Akademia Techniczno-Humanistyczna w Bielsku-Białej, Wydział Budowy Maszyn i Informatyki, specjalność: Inżyniera Mechaniczna, marcinmatuszny@wp.pl

² Prof. ATH, dr hab. inż., Akademia Techniczno-Humanistyczna w Bielsku-Białej, Wydział Budowy Maszyn i Informatyki, ipraszkievicz@ath.bielsko.pl

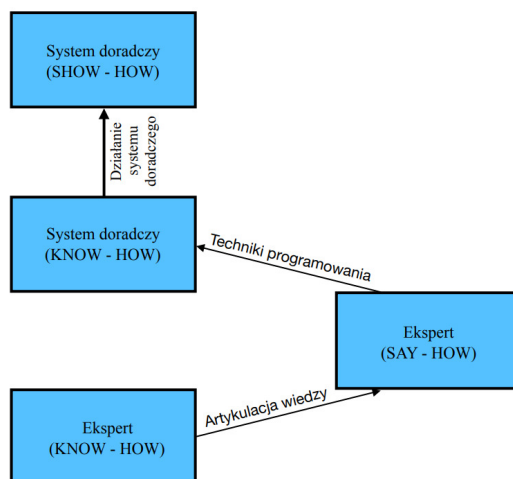
w ściśle określonym zakresie zadań z dziedziny projektowania procesów technologicznych ze zidentyfikowanych źródeł wiedzy, a także zapis pozyskanej wiedzy w sposób umożliwiający jej zastosowanie w procesie wspomagania podejmowania decyzji podczas rozwiązywania zadań dotyczących projektowania procesów technologicznych [1].

Pozyskiwanie wiedzy z punktu widzenia budowy baz wiedzy jest bardzo mocno związane z pojęciem uczenia maszynowego (ang. *machine learning*). Najbardziej ogólnie, pozyskiwanie wiedzy (ang. *knowledge acquisition*) można zdefiniować jako uczenie, w rozumieniu pozyskiwania informacji symbolicznej, połączone z nabyciem umiejętności wykorzystania wiedzy w sposób efektywny [2].

Razem z zaawansowaniem rozwoju badań na płaszczyźnie budowy systemów ekspertowych pojawiały się przeszkody, wynikające głównie z konieczności coraz bardziej rozbudowanych baz wiedzy. Niezbędnym okazało się opracowanie takich metod pozyskiwania wiedzy, które będą efektywne z uwagi na skrócenie czasu budowy i weryfikacji baz wiedzy pod względem ich niesprzeczności, zupełności oraz eliminacji nadmiarowości informacji. Najczęściej stosowanymi metodami pozyskiwania wiedzy bezpośrednio od eksperta są:

- Bezpośrednie konsultacje z ekspertem,
- Analiza i obserwacja pracy wykonywanej przez eksperta,
- Zapis wiedzy przez eksperta w specjalnie stworzonym do tego celu formularzu elektronicznym bądź papierowym;

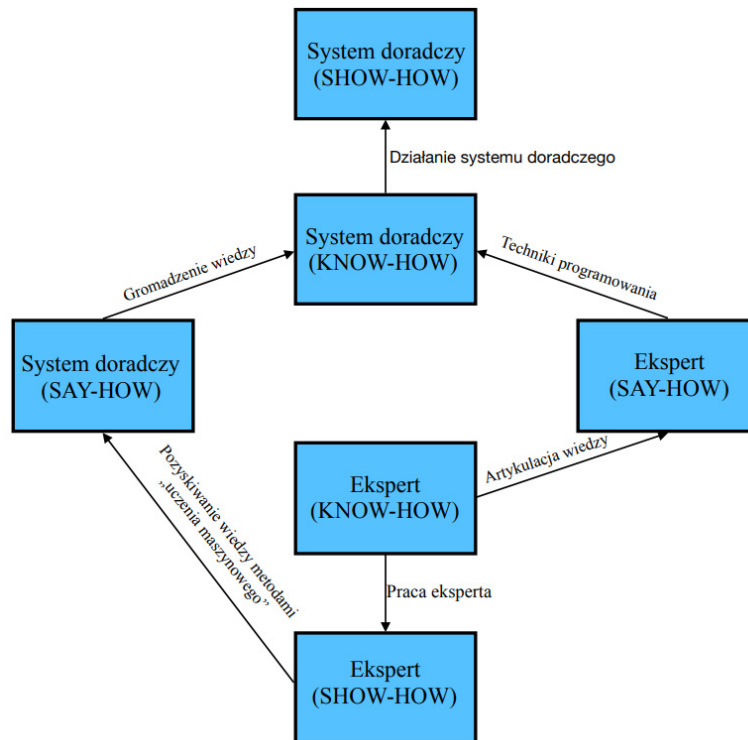
Rysunek 1 przedstawia schemat tradycyjnego procesu pozyskiwania wiedzy od eksperta. Takie podejście utrudnia weryfikację, budowę oraz kontrolę nad złożonymi bazami wiedzy ze względu na niejednokrotne problemy w artykulacji wiedzy przez tegoż eksperta oraz błędy w zapisie wiedzy.



Rysunek 1. Pozyskiwanie wiedzy w sposób "tradycyjny", opracowanie własne na podstawie [3]

Aby ograniczyć wady tradycyjnego procesu pozyskiwania wiedzy od eksperta, opracowano wiele metod, które umożliwiają komputerowe i zautomatyzowane pozyskiwanie wiedzy. W takich metodach nie wymaga się od eksperta

bezpośredniego udziału w procesie pozyskiwania wiedzy. Wiedza, która została pozyskana wyżej opisanymi metodami jest zdobyta w oparciu o wyniki prac ekspertów oraz nagromadzonych już wcześniej danych. Poniższy schemat (rys. 2) obrazuje pozyskiwanie wiedzy od eksperta w procesie wspomaganym metodami uczenia maszynowego.

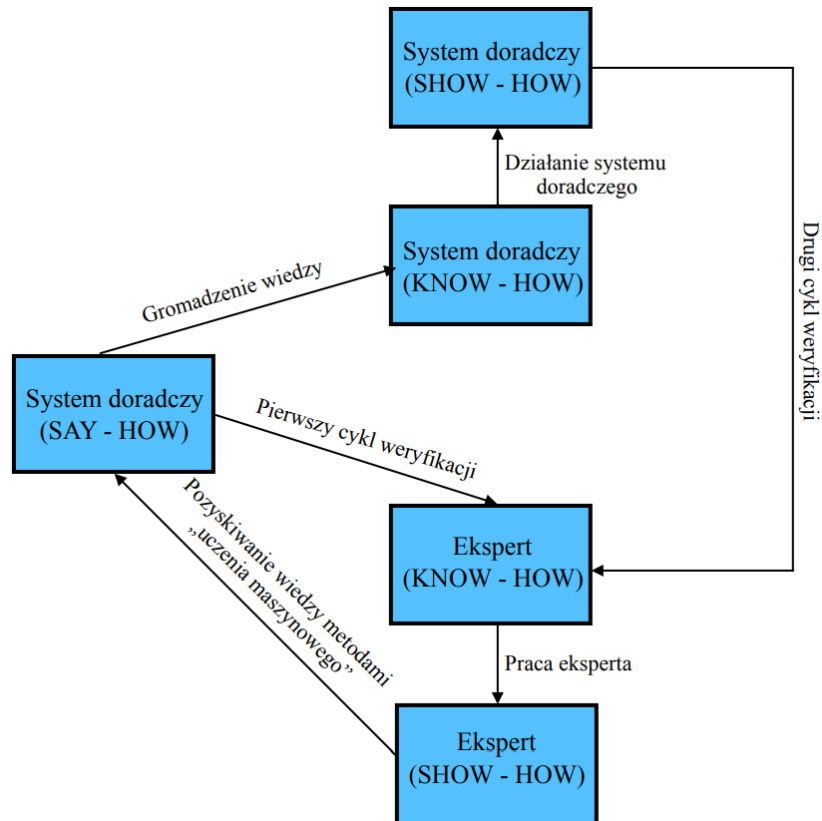


Rysunek 2. Pozyskiwanie wiedzy bezpośrednio od eksperta jako proces, który został wsparty metodami uczenia maszynowego, opracowanie własne na podstawie [3]

Schemat (rys. 3) obrazuje kolejny etap w procesach pozyskiwania wiedzy. W tym procesie ekspert nie przekazuje już swojej wiedzy, lecz dokonuje dokładnej jej weryfikacji. Wiedza została pozyskana metodami uczenia maszynowego, a zadaniem eksperta jest weryfikacja poprawności działania systemu ekspertowego, a także sprawdzenie uzyskanych z systemu wniosków i objaśnień. W przypadku stwierdzenia błędów i innych nieprawidłowości, ekspert powinien dokonać poprawek w bazie wiedzy lub nawet ponownego pozyskania tejże wiedzy na podstawie uzupełnionych i odpowiednio poprawionych danych. Przedstawiony poniżej proces jest wydajniejszy od dwóch poprzednich, ponieważ:

- pozyskanie wiedzy z użyciem metod opartych o machine learning jest znacznie szybsze,
- bazy wiedzy pozyskane w sposób zautomatyzowany zawierają znacznie mniejszą liczbę błędów lub przy optymalnym stanie, nie zawierają ich w ogóle,

- weryfikacja wiedzy pozyskanej w ten sposób przez eksperta jest dla niego zadaniem znacznie prostszym i szybszym niż artykulacja tej wiedzy.



Rysunek 3. Proces pozyskiwania wiedzy wspomagany metodami uczenia maszynowego, opracowanie własne na podstawie [3]

1.1. Metody pozyskiwania wiedzy

Podstawową ideą uczenia się jest pozyskiwanie wymaganej wiedzy z zastosowaniem kilku metod rozumowania – indukcji, dedukcji lub analogii. W szczególnym przypadku uczenie może wymagać tylko powielania informacji dostarczonych przez otoczenie lub transformację tych informacji i wydzielenie z nich pewnej istotnej dla nas części. Proces pozyskiwania wiedzy sklasyfikowany jest zależnie od wielu kryteriów. Jednym z takich kryteriów może być ilość informacji przekazanej do systemu doradczego. Można wyróżnić tutaj następujące metody pozyskiwania wiedzy:

- bezpośrednio pozyskiwanie wiedzy – nie wymaga od systemu podlegającego uczeniu wnioskowania i transformacji wiedzy, jest realizowane przykładowo przez bezpośrednie zaprogramowanie, metoda ta jest stosowana dla prostych baz wiedzy,

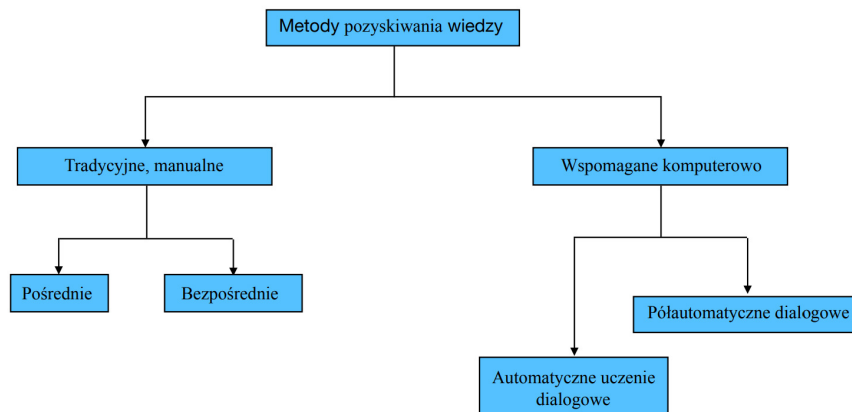
- pozyskiwanie wiedzy na podstawie instrukcji – wymaga konieczności współdziałania uczącego z uczącym się, realizowane poprzez zastosowanie odpowiednich źródeł wiedzy wskazanych przez uczącego i ich przekształcenia na język akceptowalny przez uczącego się.
- pozyskiwanie wiedzy na podstawie analogii – polega na takim przekształceniu istniejącej informacji, aby mogła być użyta do opisów faktów podobnych do tych, które już zostały zawarte w bazie wiedzy, realizowane przykładowo przez modyfikację programu komputerowego
- pozyskiwanie wiedzy na podstawie przykładów – metoda bardzo często stosowana przy konstruowaniu baz wiedzy, polega na generowaniu ogólnego opisu klas na podstawie zbioru przykładów i kontrprzykładów reprezentujących te klasy, ogólny opis otrzymywany jest na podstawie zasady indukcji.
- pozyskiwanie wiedzy na podstawie obserwacji – metoda ta wymaga większego udziału uczącego się podczas procesu uczenia, uczący może dokonywać obserwacji biernych oraz czynnych.

Spośród wielu metod indukcji wiedzy na podstawie przykładów można wyróżnić następujące metody:

- indukcja reguł za pomocą generowania pokryć [4],
- indukcja drzew decyzyjnych [5],
- indukcja reguł za zastosowaniem zbiorów przybliżonych [6].

Reasumując powyższe, metody pozyskiwania wiedzy, ze względu na poziom zaangażowania oprogramowania w pozyskiwanie wiedzy (rys. 4), można podzielić na [7]:

- manualne (ang. *manual methods*)
- półautomatyczne (ang. *semi-automatic methods*)
- automatyczne (ang. *automatic methods*)



Rysunek 4. Klasyfikacja metod pozyskiwania wiedzy na podstawie [7]

W celu zastosowania tych metod konieczne jest użycie tak zwanego atrybutowego modelu opisu przestrzeni zainteresowania, czyli określenie tzw. dziedziny, aby możliwe było pozyskiwanie wiedzy na podstawie przykładów konieczne jest wyszczególnienie:

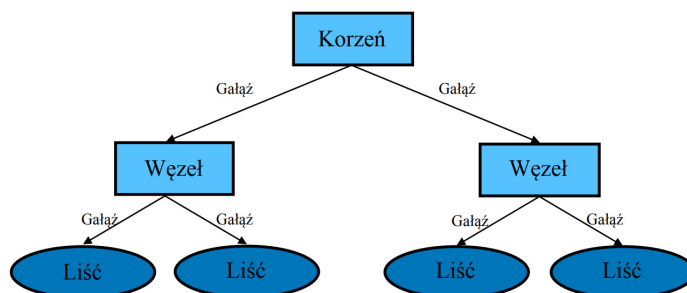
- obiektów – zjawisk, problemów, które podlegają klasyfikacji,
 - atrybutów opisujących dane obiekty, wartości jakie przyjmą poszczególne atrybuty.
 Wartości atrybutów dla wybranej grupy obiektów tworzą zbiór uczący, a przykłady opisujące obiekty nazywa się przykładami uczącymi.

Mając do dyspozycji powyższe informacje, można rozpocząć pracę nad indukcją drzew decyzyjnych, z których następnie tworzone są reguły. Algorytm Quinlana, który jest bardzo pomocny przy tym procesie stosowany jest wówczas, gdy mamy do dyspozycji dane, na podstawie których możliwe są do wygenerowania reguły. Budowanie drzew decyzyjnych na podstawie danych może prowadzić do bardzo rozbudowanych drzew, a co za tym idzie do skomplikowanych, rozbudowanych reguł, nie dających efektów w procesie wnioskowania, gdyż często w oparciu o dane nie widać, że z obecności jakiejś cechy lub jej braku wynika natychmiastowy wniosek. Dzięki zastosowaniu oprogramowania opartego o algorytm Quinlana można określić kolejność atrybutów dobieranych przy budowie drzewa decyzyjnego [8].

2. Struktura drzewa decyzyjnego

Aby możliwa była poprawna indukcja drzewa decyzyjnego, bardzo ważnym jest zapoznanie się ze strukturą drzewa decyzyjnego (rys. 5). Zgodnie z formalną definicją drzewa decyzyjnego - jest to struktura drzewiasta, której każdy węzeł odpowiada przeprowadzeniu pewnego testu na wartości jednego atrybutu, zaś każdy liść zawiera decyzję o klasyfikacji przykładu. Z poszczególnych węzłów wychodzi tyle gałęzi, ile jest możliwych wyników testu odpowiadających tym węzłom. Każda z tych gałęzi prowadzi do poddrzewa (węzła) służącego do klasyfikacji tych obiektów, dla których ten test ma określony wynik.

Drzewa decyzyjne konstruowane są przez rekurencyjne dzielenie danych treningowych na podzbiory na podstawie tego, w jaki sposób ten podział wpłynie na stany zmiennej wyjściowej [9].



Rysunek 5. Struktura drzewa decyzyjnego przedstawiona w formie graficznej prezentacji z nazewnictwem poszczególnych elementów składowych drzewa, opracowanie własne na podstawie [10]

Drzewa decyzyjne są strukturami przeznaczonymi do rozwiązywania problemów klasyfikacji. Wyżej opisane węzły drzewa decyzyjnego reprezentują atrybuty rozważanego problemu. Krawędzie drzewa są związane ze skończonymi zbiorami wartości atrybutów. Drzewo decyzyjne powinno budowane być poczynając od

najbardziej znaczącego atrybutu. Następnie są wykorzystywane pozostałe atrybuty na niższych poziomach drzewa. Wybór atrybutu oparty jest również na jego zdolnościach do klasyfikacji. Miara istotności atrybutu ma charakter entropii. Po zbudowaniu drzewa, każdy nowy obiekt może być klasyfikowany poprzez przejście przez drzewo w dół – od korzenia (wierzchołka) do liścia (węzła końcowego) [11].


3. Wykorzystanie oprogramowania dedykowanego do indukcji drzew decyzyjnych

W tej części pracy przedstawiono praktyczne wykorzystanie oprogramowania DeTreeX 4.0 dedykowanego do indukcji drzew decyzyjnych wchodzącego w skład pakietu sztucznej inteligencji Sphinx 4.0. Aplikacja DeTreeX jest narzędziem służącym przede wszystkim do wspomagania procesu pozyskiwania wiedzy. Dzięki zastosowanej indukcyjnej metodzie uczenia maszynowego możliwe jest budowanie drzew decyzyjnych i zapis tych drzew w postaci reguł, gdzie warto zaznaczyć, że reguły są najczęściej stosowana metoda reprezentacji wiedzy w bazach wiedzy systemów ekspertowych.

DeTreeX 4.0 jest niezależny dziedzinowo i może zostać wykorzystany nie tylko do budowy drzew decyzyjnych w procesach produkcyjnych, ale także w innych dziedzinach między innymi w ekonomii czy medycynie. Aplikacja może być stosowana tam, gdzie pojawia się problem podejmowania decyzji, klasyfikacji obiektów, szybkiej weryfikacji pozyskanych reguł czy szybkiego pozyskania reguł decyzyjnych ze zbioru przykładów uczących.

Aby możliwym była indukcja drzewa decyzyjnego z wykorzystaniem oprogramowania oparto się na zbiorze uczącym, który dotyczy procesu produkcji szkła, w opisanym przykładzie zmienną wyjściową jest rodzaj produkowanego szkła, natomiast zmiennymi wejściowymi są: współczynnik załamania, oraz zawartość poszczególnych pierwiastków w produkowanym szkłe: sód, magnez, aluminium, krzem, potas, wapń, bar i żelazo.

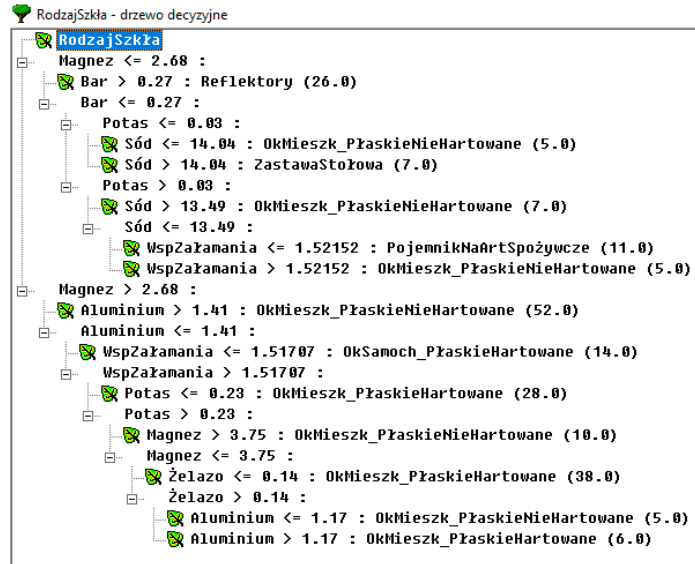
W oparciu o zbiór uczący składający się z 214 przykładów wygenerowano dziedzinę oraz przedstawiono poniżej atrybuty wraz z ich wartościami podanymi w formie przedziałów (rys. 6).

 Dziedzina problemu

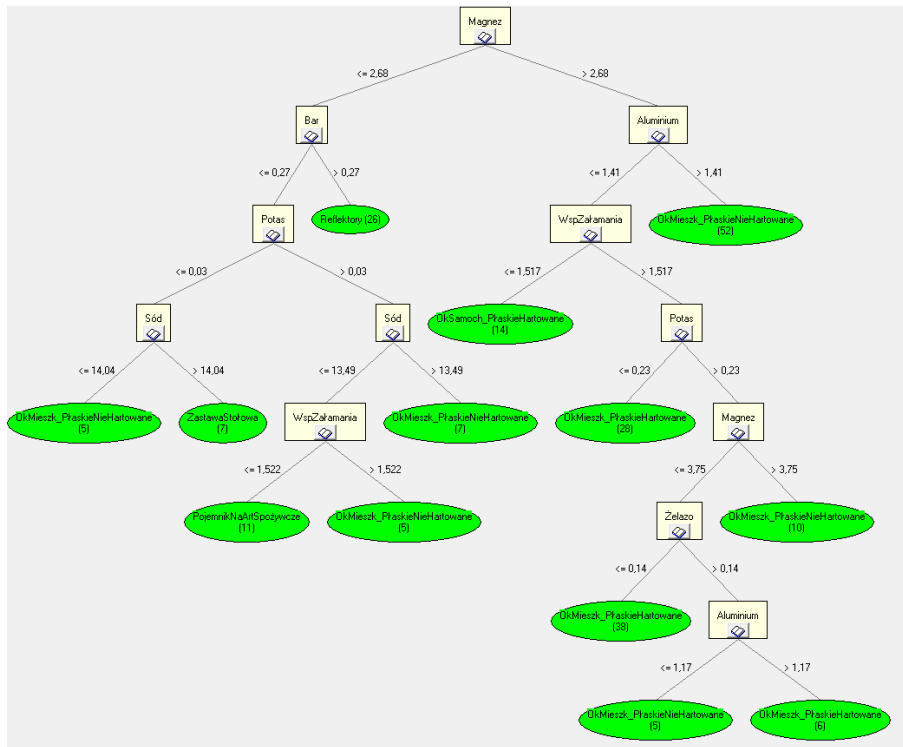
WspZalamania	Sód	Magnez	Aluminium	Krzem	Potas	Wapń	Bar	Żelazo	RodzajSzkła
min. 1.51115	min. 10.73	min. 0.00	min. 0.29	min. 69.61	min. 0.00	min. 5.43	min. 0.00	min. 0.00	OkMieszk_PlaskieHartowane
max. 1.53393	max. 17.38	max. 4.49	max. 3.50	max. 75.41	max. 6.21	max. 16.19	max. 3.15	max. 0.51	OkMieszk_PlaskieNieHartowane
									OkSamoch_PlaskieHartowane
									PojemnikNaArtSpozywcze
									ZastawaStolowa
									Reflektory

Rysunek 6 Dziedzina problemu przedstawiona w formie zrzutu ekranu z oprogramowania DeTreeX 4.0

Następnie, korzystając z opcji generowania drzewa decyzyjnego, przyjęto minimalną liczbę przykładów tworzących liść drzewa jako 5 oraz wybrano opcję przycinania drzewa o 25% (przy założeniu programowym, że będzie dostępna taka możliwość). Po ustawieniu i zatwierdzeniu wyżej wymienionych parametrów wygenerowano drzewo. Wygenerowane drzewo jest czytelne, a efekt prac, w formie drzewa graficznego i tekstowego zaprezentowano poniżej, odpowiednio na rysunku 7 i 8.



Rysunek 7. Drzewo decyzyjne wygenerowane przy użyciu oprogramowania DeTree 4.0 przedstawione w formie tekstowej



Rysunek 8. Drzewo decyzyjne wygenerowane przy użyciu oprogramowania DeTree 4.0 przedstawione w formie graficznej

4. Wnioski

W pracy przedstawiono metody pozyskiwania wiedzy z uwzględnieniem metody indukcji drzew decyzyjnych, którą zaprezentowano na praktycznym przykładzie z wykorzystaniem dedykowanego oprogramowania Sphinx 4.0. Oprogramowanie Sphinx 4.0 pozwala w relatywnie prosty sposób wygenerować drzewa decyzyjne, a także w oparciu o czytelny interfejs użytkownika pozyskać z nich poszukiwaną wiedzę w postaci reguł w oparciu o zbiór uczący. Pozyskiwanie wiedzy za pomocą drzew decyzyjnych jest bardzo efektywne dla problemów decyzyjnych w przedsiębiorstwie produkcyjnym, ponieważ posiadają one heurystyczny charakter rozwiązań. W przypadku takich problemów stosowanie systemów eksperckich w wybranych obszarach winno być celowe.

Przesłanką do opracowania metody budowy bazy wiedzy dla procesów produkcyjnych, dla obszaru technologicznego przygotowania produkcji jest symboliczna i obiektowa reprezentacja danych opisujących te procesy oraz związki zachodzące między nimi, które mogą być zapisane w postaci reguł przy zastosowaniu odpowiedniej metody pozyskiwania wiedzy oraz poprawnie przeprowadzonej indukcji drzew decyzyjnych.

Warto również zaznaczyć, że poprawnie opracowane schematy procesu przetwarzania wiedzy technologicznej, poczynwszy od jej pozyskiwania, poprzez reprezentację, a skończywszy na utworzeniu bazy wiedzy, umożliwiają przygotowanie i zapis niezbędnej wiedzy w systemie eksperckim, pozwalającej na optymalizację kosztów produkcji oraz optymalizację wszystkich parametrów procesu produkcyjnego w przedsiębiorstwie.

LITERATURA

1. KNOSALA R. i zespół: Zastosowania metod sztucznej inteligencji w inżynierii produkcji. Wydawnictwo Naukowo-Techniczne, Warszawa 2002.
2. CHOLEWA W., PEDRYCZ W.: Systemy doradcze, skrypt uczelniany Politechniki Śląskiej nr 1447, Gliwice 1987.
3. MICHIE D.: Current Developments in Experts Systems. Applications of Expert Systems, Turing Institute Press in association with Addison-Wesley Publishing Company, 1987.
4. MICHALSKI R. S.: A Theory and Methodology of Inductive Learning, Artificial Intelligence 20, pp. 111-161, 1983.
5. QUINLAN J. R.: C4.5 Program for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.
6. OHM A., KOMOROWSKI J., SKOWRON A., SYNAK P.: The Design and Implementation of a Knowledge Discovery Toolkit Based on Rough Sets – The Rosseta System. Rough Sets in Knowledge Discovery, Physica Verlag, 1998.
7. JAGIELSKI J.: Inżynieria wiedzy. Oficyna Wydawnicza Uniwersytetu Zielonogórskiego, Zielona Góra 2005.
8. KWIATKOWSKA A. M.: Systemy wspomaganie decyzji. Jak korzystać z wiedzy i informacji w praktyce. Wydawnictwo Naukowe PWN SA, Warszawa 2007.

9. SZELIGA M.: Data science I uczenie maszynowe. Wydawnictwo Naukowe PWN SA, Warszawa 2017.
10. POPOVA O., POPOV B., KARANDEY V., GERASHCHENKO A.: Entropy and Algorithm of Obtaining Decision Trees in a Way Approximated to the Natural Intelligence. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 13(3), 2019.
11. PEDRYCZ W., SOSNOWSKI Z. A.: Drzewa decyzyjne z rozmytą granulacją wiedzy. *Inżynieria Wiedzy i Systemy Ekspertowe*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław 2000.