



WYKRYWANIE PRZYCZYŃ I PRZEWIDYWANIE PROBLEMÓW Z JAKOŚCIĄ NA PRZYKŁADZIE PRZEMYSŁU POLIGRAFICZNEGO

Tomasz Demski, StatSoft Polska Sp. z o.o.

Artykuł poświęcony jest tworzeniu modelu procesu produkcyjnego w celu wykrycia przyczyn powstawania problemów z jakością i przewidywania, czy dla konkretnego cyklu procesu istnieje zagrożenie wystąpienia wad. Przedstawiony przykład dotyczy druku rotograviurowego, w którym z niewyjaśnionych przyczyn co jakiś czas pojawiały się niepożądane pasy na wydrukach (tzw. *banding*). Za pomocą narzędzi data mining zbudujemy model wskazujący czynniki wpływające na zagrożenie wystąpienia wady oraz umożliwiający przewidywanie wystąpienia takiego zjawiska. Na koniec uzyskane rozwiązanie zostanie wdrożone w *STATISTICA Enterprise*.

Wprowadzenie

Współczesne procesy wytwarzania są bardzo często skomplikowane, mają bardzo wiele cech. Zazwyczaj w toku procesu zbieramy mnóstwo danych: ustawień procesu, właściwości surowców oraz parametrów, takich jak temperatura i ciśnienie. Zdarza się, że od czasu do czasu w procesie występują problemy z jego jakością, albo że następuje trwałe obniżenie wydajności procesu. Zastosowanie technik analizy danych ułatwia szybkie wykrycie przyczyny tego typu problemów. Wprowadzenie do zagadnienia data mining i przykład zastosowania go do wykrywania przyczyn obniżenia jakości w przemyśle półprzewodnikowym znajduje się w artykule [1].

Narzędzia data mining możemy wykorzystywać do rozwiązywania problemów z jakością procesów na kilka sposobów, np.:

- ◆ Wykrycie sekwencji zdarzeń powodujących obniżenie jakości lub wydajności. Badamy, które zmienne wpływają na wydajność lub wybraną miarę jakości. Pod uwagę bierzemy wszystkie zmienne informujące o wystąpieniu różnych zdarzeń (np. użycie konkretnego narzędzia), ustawieniach procesu i jego przebiegu. W ten sposób możemy się dowiedzieć, że np. jeśli w toku przygotowania wafla półprzewodnikowego użyto dwóch konkretnych narzędzi, to uzyskano mniej poprawnie działających układów (zob. [1]).
- ◆ Ostrzeżenie o zmianie w procesie. W tym wypadku tworzymy model, który dobrze opisuje nasz proces. Jeśli proces zmieni się, to trafność modelu gwałtownie spadnie.



Dzięki temu możemy szybko wykryć zakłócenia procesu, nawet gdy jest on bardzo skomplikowany.

◆ Przewidywanie wyniku procesu.

My zajmiemy się problemem tzw. *bandingu*. Występuje on przy wysokonakładowym druku techniką rotograwiową. W pewnych sytuacjach, przy dużej szybkości druku na wydrukach pojawiały się pasy, powodujące, że materiały nadawały się tylko do wyrzucenia. Niestety tradycyjnymi metodami nie udało się, pomimo zaangażowania fachowców, określić przyczyny pojawiania się uszkodzeń na wydrukach i do ich wykrycia zdecydowano się zastosować narzędzia data mining. Projekt data mining zakończył się powodzeniem i pozwolił zasadniczo zmniejszyć częstość występowania problemu. Dokładniejszy opis oryginalnego projektu znajduje się w podręczniku [2].

W tym przykładzie zbudujemy model przewidujący wystąpienie pasów na wydrukach na podstawie danych udostępnionych w składnicy danych *UCI Machine Learning* [3]. Jako narzędzie udostępniania danych do analizy wykorzystamy *STATISTICA Enterprise*, a do budowy modelu użyjemy *Przepisów STATISTICA Data Miner*.

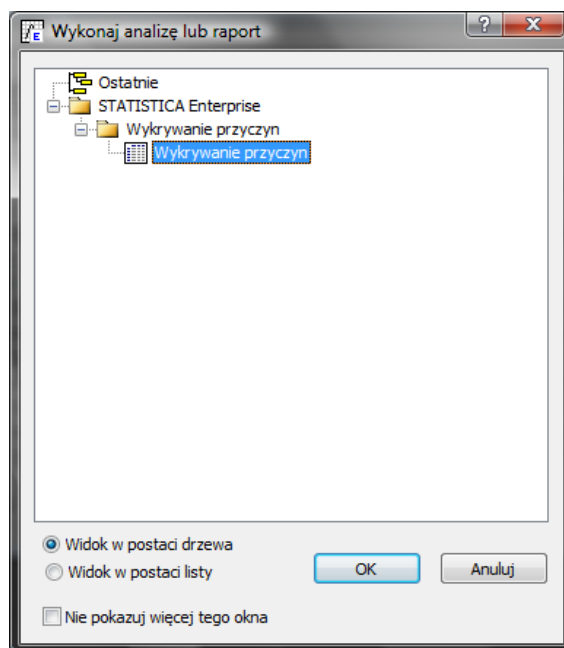
Podstawowe informacje o Przepisach *STATISTICA Data Miner*

Narzędzia zgłębiania danych (*data mining*) rozwijane są od wielu lat i obecnie dostępnych jest wiele dojrzałych metod dostosowanych do rozmaitych zadań i wymagań dotyczących stosowania uzyskiwanych rozwiązań. Równocześnie postęp w szybkości działania komputerów i pojemności pamięci masowych powoduje, że coraz mniejszą przeszkodę w analizie stanowi wielkość danych. W związku z tym coraz ważniejsze jest ułatwienie wykonywania **całego** procesu prowadzącego od surowych danych do wiedzy, z uwzględnieniem przygotowania i oczyszczenia danych oraz stosowania modeli dla nowych przypadków. Właśnie z myślą o ułatwieniu wykonywania analiz w praktyce przygotowano *Przepisy STATISTICA Data Miner*.

Przepisy STATISTICA Data Miner umożliwiają rozwiązywanie zadań ukierunkowanego data mining poprzez wykonanie określonej sekwencji operacji. Użytkownik jest prowadzony przez całą analizę: od wskazania danych, poprzez ich sprawdzenie, oczyszczenie i przekształcenie, zbudowanie modelu i zastosowanie go dla nowych przypadków. Na koniec można utworzyć raport podsumowujący wszystkie wykonane działania.

Tworzenie modelu

Zaczynamy od pobrania danych. W *STATISTICA Enterprise* istnieje już konfiguracja analizy wykonującą pobranie odpowiednich danych (tworzenie konfiguracji danych i analiz przedstawiono w [4]). Po uruchomieniu programu w oknie *Wykonaj analizę lub raport* (widocznym poniżej) wskazujemy konfigurację *Wykrywanie przyczyn* i klikamy *OK*.




Rys. 1. Pobieranie danych do modelowania.

Program wczyta dane z bazy danych i wyświetli je w arkuszu *STATISTICA*. Zauważmy, że w praktyce uzyskanie danych do tworzenia i stosowania modelu często jest złożone. Dlatego przygotowanie i zapisanie szablonu pobierania danych daje bardzo duże korzyści: analityk nie musi wiedzieć, jak wydobyć dane ze źródłowych systemów bazodanowych, a zamiast wykonywać wiele operacji ręcznie, po prostu wybiera potrzebną mu konfigurację i klika *OK*.

Fragment arkusza z pobranymi danymi widzimy na rys. 2. Łącznie w arkuszu mamy 38 zmiennych i 540 przypadków. Informacja, czy wystąpiły uszkodzenia wydruków znajduje się w zmiennej *band type*. Zmienne *timestamp*, *cylinder number* i *job number* stanowią identyfikator poszczególnych procesów (nieomal dla każdego przypadku przyjmują inną wartość) i dlatego nie wykorzystamy ich w analizie. W modelowaniu pominiemy również zmienną *customer*, tj. identyfikator klienta. W skład danych wchodzi zmienna *test*, którą wykorzystamy do podziału danych na przeznaczone do tworzenia modelu i do jego oceny. Ostatecznie w budowie modelu jako wejścia (zmienne niezależne) wykorzystamy 32 zmienne.

Przed właściwą analizą konieczne jest oczyszczenie danych, np. usunięcie lub wypełnienie braków danych, odrzucenie zbędnych zmiennych itp. W naszym przypadku wykonamy je za pomocą automatycznych procedur zawartych w *Przepisach Data Miner*.

Po wczytaniu danych uruchamiamy *Przepisy Data Miner*. W tym celu przechodzimy na kartę *Data Mining* i naciskamy przycisk  na wstążce. Na ekranie pojawi się okno *Przepisów*, w którym klikamy przycisk *Nowy*.



STATISTICA - [Dane: Wykrywanie przyczyn.sta (38 zm., * 540 prz.)]

Podstawowe Edycja Widok Wstaw Format Statystyka Data Mining Wykresy Narzędzia Dane Korporacyjne Pomoc

Statystyki podstawowe wieloraka Regresja ANOVA Nieparametryczne Dopasowanie rozkładu Więcej rozkładów

Modely zaawansowane Sieci neuronowe Karty kontrolne Analiza procesu

Wielowymiarowe PLS, PCA, ... Wielowymiarowe DOE

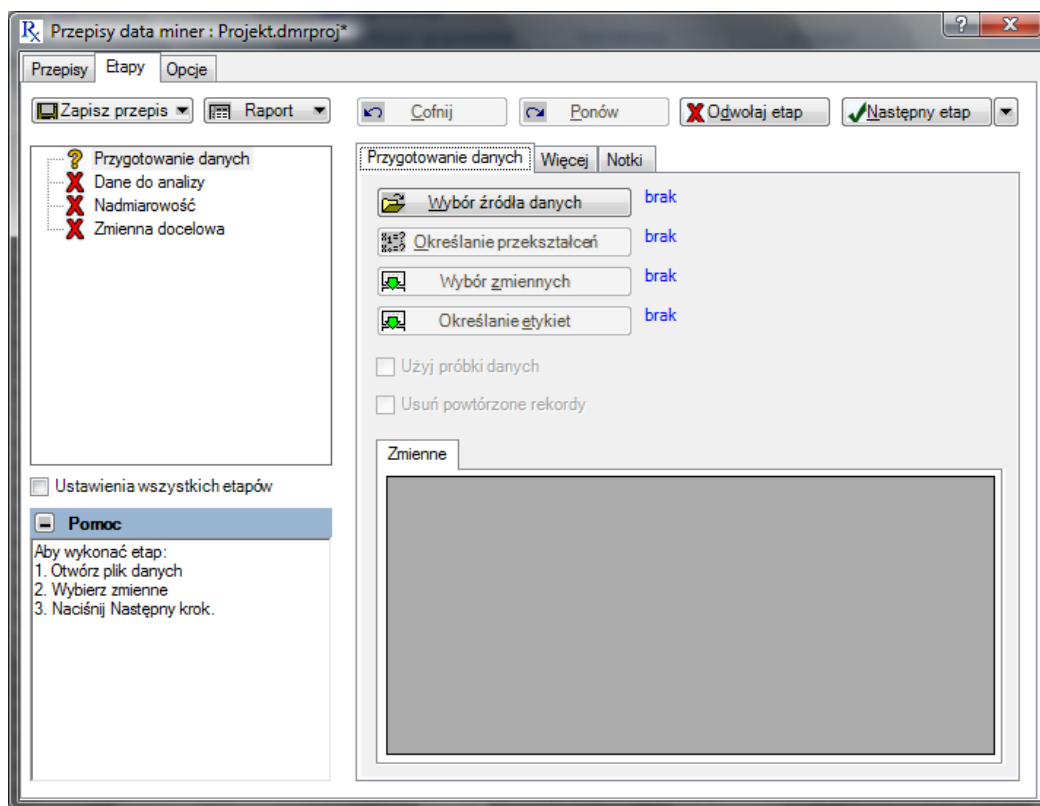
Analiza mocy testu Wariacja Predykcyjne Szóst Sigm

STATISTICA VB Grupami Kalkulatory Statystyki bloku

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	timestamp	cylinder number	customer	job number	grain screened	proof on ctd ink	blade mfg	paper type	ink type	direct steam	solvent type	type on cylinder	press type	press cy
1	19910108	X126	TVGUIDE	25503	YES	YES	BENTON	UNCOATED	UNCOATED	NO	LINE	YES	Motter94	821 TAI
2	19910109	X266	TVGUIDE	25503	YES	YES	BENTON	UNCOATED	UNCOATED	NO	LINE	YES	Motter94	821 TAI
3	19910104	B7	MODMAT	47201	YES	YES	BENTON	UNCOATED	COATED	NO	LINE	YES	WoodHoe70	815 CA
4	19910104	T133	MASSEY	39039	YES	YES	BENTON	UNCOATED	UNCOATED	NO	LINE	YES	WoodHoe70	816 CA
5	19910111	J34	KMART	37351	NO	YES	BENTON	UNCOATED	COATED	NO	LINE	YES	WoodHoe70	816 TAI
6	19910104	T218	MASSEY	38039	YES	YES	BENTON	UNCOATED	UNCOATED	NO	LINE	YES	WoodHoe70	816 CA
7	19910111	X249	ROSES	35751	NO	YES	BENTON	COATED	COATED	NO	LINE	YES	Motter94	827 TAI
8	19910111	X788	ROSES	35751	NO	YES	BENTON	COATED	COATED	NO	LINE	YES	Motter94	827 TAI
9	19910112	M372	MODMAT	47201	YES	YES	BENTON	UNCOATED	UNCOATED	NO	XYLOL	YES	Albert70	802 CA
10	19910114	I320	CHILDCRAFT	37000	YES	YES	BENTON	UNCOATED	UNCOATED	NO	LINE	YES	WoodHoe70	815 CA
11	19910114	I337	CHILDCRAFT	37000	YES	YES	BENTON	UNCOATED	UNCOATED	NO	LINE	YES	WoodHoe70	815 CA
12	19910111	X352	HANOVRRHOUSE	35539	YES	YES	BENTON	UNCOATED	UNCOATED	NO	LINE	YES	Motter94	821 TAI
13	19910117	X67	HANOVRRHOUSE	35539	YES	YES	BENTON	UNCOATED	UNCOATED	NO	LINE	YES	Motter94	821 TAI
14	19910125	X817	GUIDEPOSTS	23052	YES	YES	BENTON	UNCOATED	UNCOATED	NO	LINE	YES	Motter94	821 TAI
15	19910117	X273	HANOVRRHOUSE	35539	YES	YES	BENTON	UNCOATED	UNCOATED	NO	LINE	YES	Motter94	821 TAI
16	19910103	F108	MODMAT	47201	?	YES	BENTON	COATED	COATED	NO	LINE	NO	Motter70	813 CA
17	19910129	F237	HOMESHOP	38064	NO	YES	BENTON	COATED	COATED	NO	LINE	NO	Motter70	813 CA
18	19910129	F267	HOMESHOP	38064	NO	YES	BENTON	COATED	COATED	NO	LINE	NO	Motter70	813 CA
19	19910123	S21	USCAV	35521	YES	YES	BENTON	UNCOATED	UNCOATED	NO	LINE	YES	Motter70	813 CA
20	19910123	F492	USCAV	35521	YES	YES	BENTON	UNCOATED	UNCOATED	NO	LINE	YES	Motter70	813 CA
21	19910130	X799	COLORTIL	37571	NO	YES	BENTON	COATED	COATED	NO	NAPHTHA	NO	Motter94	824 TAI
22	19910130	X823	COLORTIL	35751	NO	YES	BENTON	COATED	COATED	NO	NAPHTHA	YES	Motter94	824 TAI
23	19910117	X163	WARDS	37340	NO	YES	BENTON	COATED	COATED	NO	LINE	YES	Motter94	824 TAI
24	19910127	O4	TARGET	35816	NO	YES	BENTON	COATED	COATED	NO	LINE	YES	Motter94	827 SP
25	19910131	E84	KMART	37352	NO	YES	BENTON	COATED	COATED	NO	LINE	YES	Motter94	827 TAI
26	19910201	E81	KMART	37352	NO	YES	BENTON	COATED	COATED	NO	LINE	YES	Motter94	827 TAI
27	19910130	E72	KMART	37352	NO	YES	BENTON	UNCOATED	COATED	NO	LINE	YES	Motter94	827 TAI
28	19910130	E310	KMART	37352	NO	YES	BENTON	UNCOATED	COATED	NO	LINE	YES	Motter94	827 TAI
29	19910201	X29	WOOLWORTH	37080	NO	YES	BENTON	COATED	COATED	NO	LINE	NO	Motter94	828 TAI

Aby uzyskać pomoc naciśnij F1. Wykrywanie pr P1,Z1 19910108 Selekcja:Nie Waga:Nie CAP NUM REC

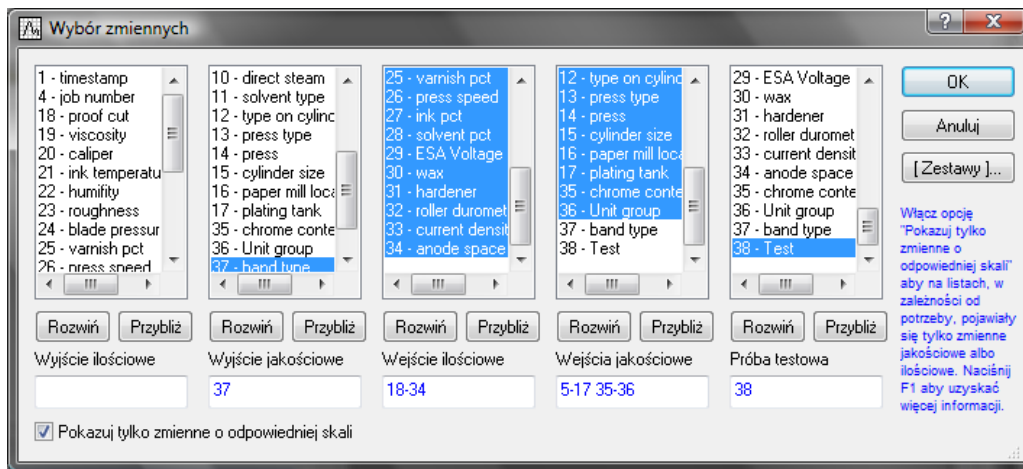
Rys. 2. Dane pobrane do STATISTICA.



Rys. 3. Okno Przepisów Data Miner.

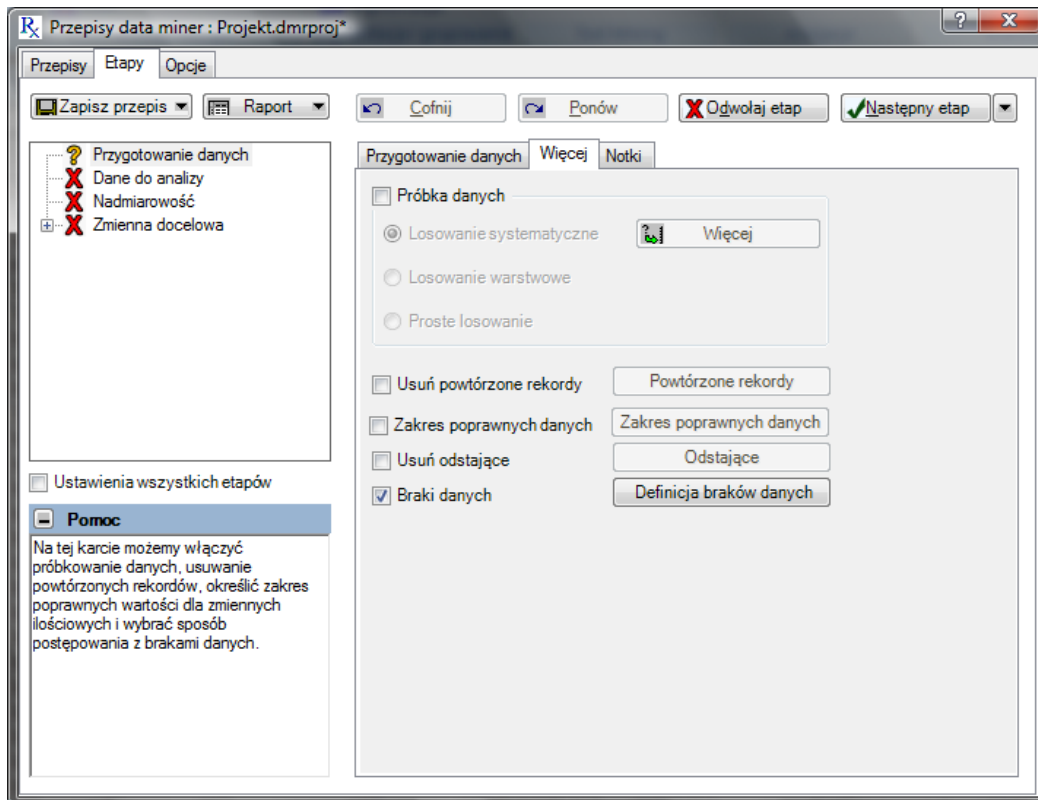


Tworzenie modelu zaczynamy od wskazania źródła danych: klikamy przycisk *Wybór źródła danych* i wskazujemy arkusz *Wykrywanie przyczyn*. Możemy teraz wybrać zmienne do modelowania, w tym celu naciskamy przycisk *Wybór zmiennych*.



Rys. 4. Wybór zmiennych.

Jako wyjście jakościowe (zmienną zależną) wskazujemy *band type*. Jako wejścia ilościowe wskazujemy wszystkie zmienne widoczne na tej liście oprócz dwóch pierwszych: *timestamp* i *job number*. Podobnie postępujemy przy wyborze wejść jakościowych, pomijając przy wyborze zmienne *cylinder number* i *customer*. Na koniec na liście *Próba testowa* wskazujemy zmienną *Test*.

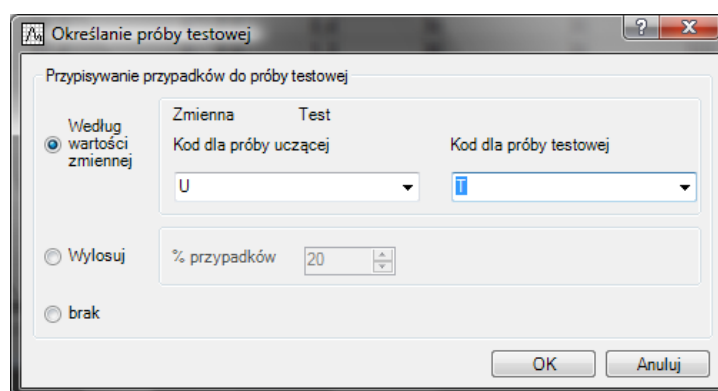


Rys. 5. Czyszczenie danych.

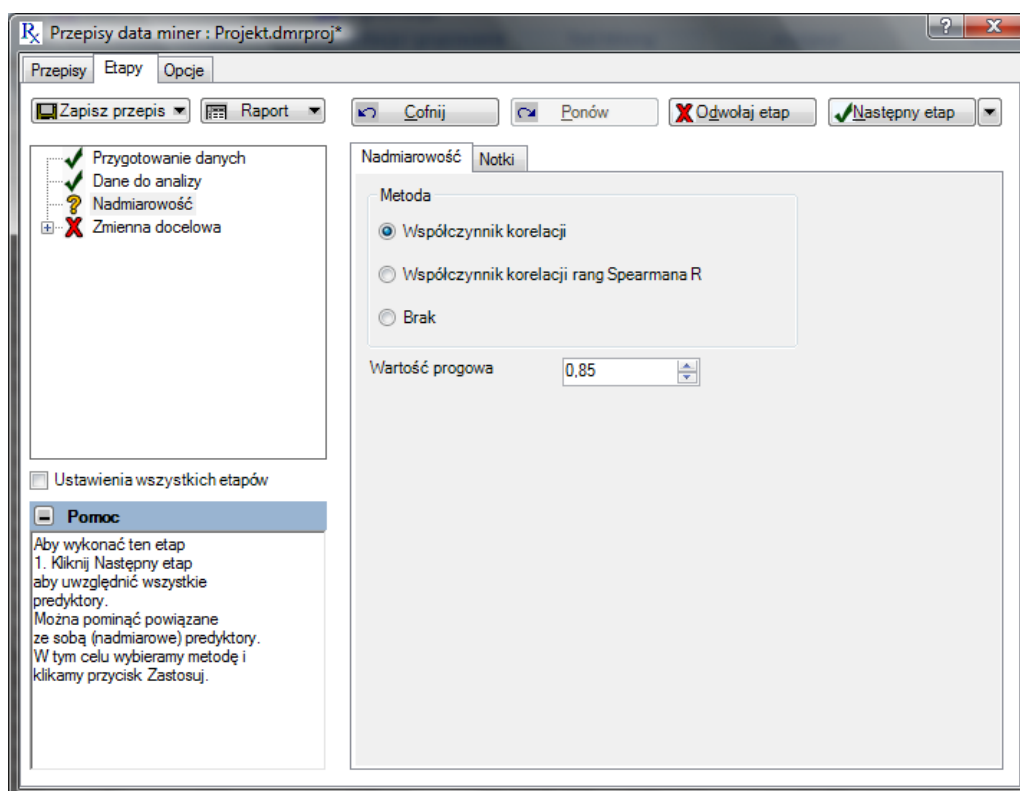


Na karcie *Więcej* (zob. rys. 5) możemy określić sposób czyszczenia danych. Do dyspozycji mamy losowe próbkowanie, usuwanie powtórzonych rekordów, wykrywanie obserwacji poza dopuszczalnym zakresem, obsługę nietypowych obserwacji i braków danych. W naszym projekcie wykorzystamy tylko obsługę braków danych.

Do kolejnego etapu przechodzimy, naciskając przycisk *Następny etap*. Program sprawdzi dane i, jeśli nie wykryje żadnych problemów, przejdziemy od kolejnego etapu, w którym określamy podział na próbę uczącą i testową. Próba ucząca zostanie użyta do znalezienia modelu, a testowa do oceny jego działania. Na karcie *Dane do analizy* klikamy przycisk *Określanie próby testowej*. Na ekranie otworzy się przedstawione niżej okno, w którym wybieramy podział na próby wg wartości zmiennej *Test*, przy czym jako kod dla próby uczącej wybieramy *U*, a dla próby testowej *T* (tak jak na rys. 6)



Rys. 6. Określanie próby testowej.



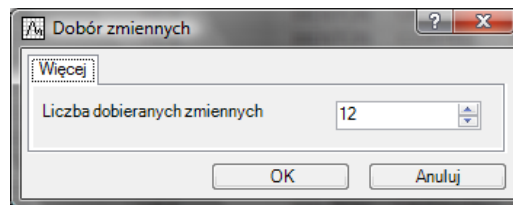
Rys. 7. Nadmiarowość.



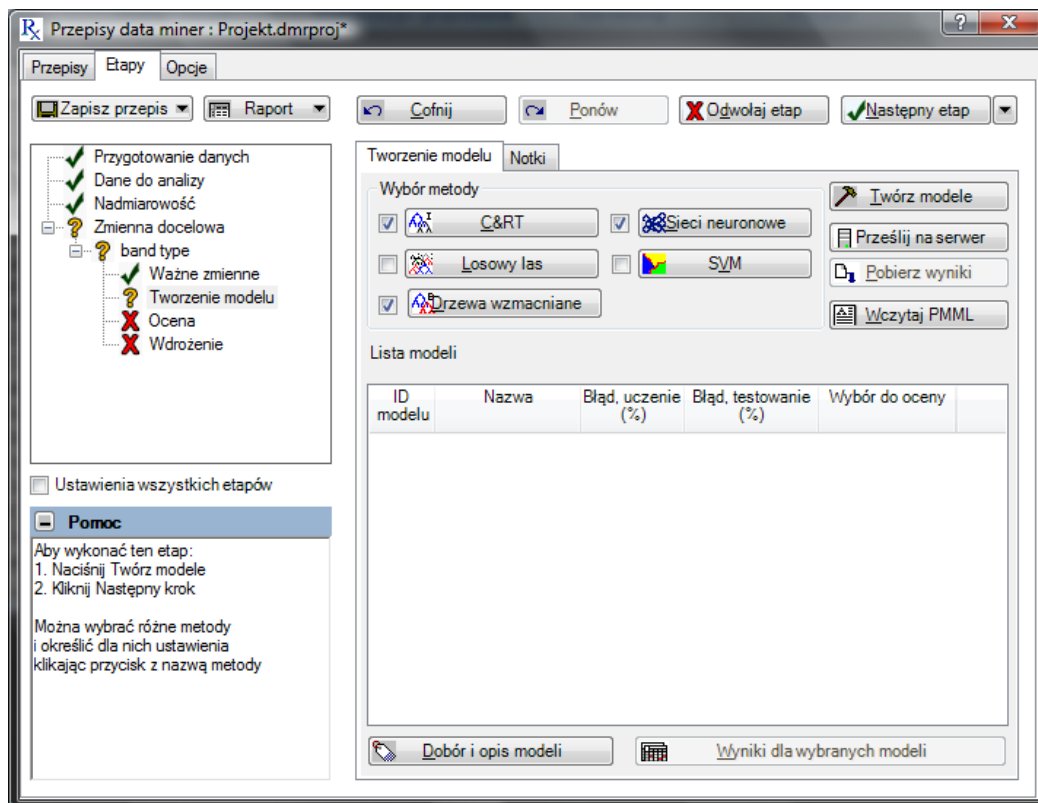
Po naciśnięciu przycisku *Następny etap* przechodzimy do kolejnej fazy modelowania: usuwania zmiennych nadmiarowych. W praktycznych zastosowaniach dosyć często zdarza się, że dane zawierają zmienne przenoszące tę samą informację. Na etapie *Nadmiarowość* możemy wyeliminować takie zmienne z modelu. W naszym przypadku jako zbędne uznamy zmienne, których współczynnik korelacji wynosi co najmniej 0,85.

Po kliknięciu *OK* program znajdzie pary zmiennych o współczynniku korelacji co najmniej 0,85 i zaproponuje usunięcie zbędnych zmiennych. W naszym przypadku nadmiarowe zmienne to *varnish pct* i *ink pct*, a ich współczynnik korelacji wynosi -0,86. Zgodnie z podpowiedzią programu w dalszym modelowaniu pominiemy zmienną *varnish pct*.

Kolejny etap to odrzucenie zmiennych niewpływających na występowanie uszkodzeń na wydrukach. W naszym przykładzie mamy stosunkowo niewiele zmiennych, które opisują proces i mogą wpływać na jego wynik. Jednak w wielu zastosowaniach, zwłaszcza w przemyśle zdarza się, że mamy dosłownie tysiące zmiennych oddziałujących na wyjście, z których tylko kilka jest naprawdę istotnych.



Rys. 8. Dobór ważnych zmiennych.



Rys. 9. Tworzenie modelu.



Do odsiania niepotrzebnych zmiennych zastosujemy *Szybki dobór zmiennych*, przy czym do dalszej analizy wybierzmy 12 zmiennych, najsilniej wpływających na wystąpienie uszkodzeń na wydrukach. Klikamy przycisk *Szybki dobór zmiennych* i ustawiamy wybór 12 najlepszych predyktorów.

Następny etap *Przepisu* to tworzenie modeli. Najpierw utworzymy modele przy domyślnych ustawieniach. Po kliknięciu *Twórz model* program znajdzie najlepsze modele metodami drzew decyzyjnych (C&RT), drzewami wzmacnianymi (*boosted trees*) oraz różnymi architekturami sieci neuronowych (opis metod można znaleźć w [5] i [6]).

Program znajdzie najlepsze modele i wyznaczy stopę błędów w próbach uczącej i testowej. Obliczenia najdłużej trwają dla sieci neuronowych, warto jednak zauważyć, że w tym wypadku sprawdzane jest kilka architektur sieci neuronowych (domyślnie 5) i wybierana jest najlepsza z nich.

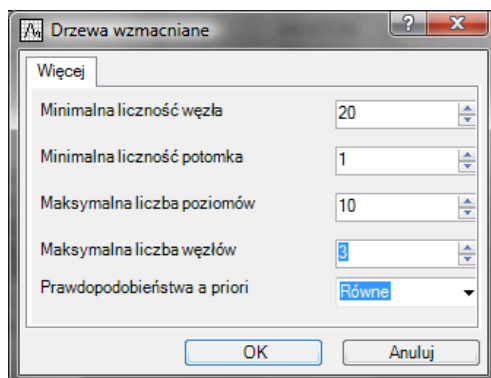
W tabeli poniżej znajduje się zestawienie stóp błędów dla różnych metod. Model możemy uznać za poprawny, gdy trafność przewidywania w obu próbach jest podobna, a najlepszy model to ten, który ma najmniejszy udział błędnych przewidywań w próbie testowej. W naszym przypadku najlepszy wydaje się model sieci neuronowej. Bardzo podobna jest trafność przewidywań drzew klasyfikacyjnych C&RT. Z kolei drzewa wzmacniane mają zdecydowanie większą stopę błędów w próbie testowej, innymi słowy są przeuczone.

Metoda	Stopa błędów (%)	
	Próba ucząca	Próba testowa
C&RT	21,50	27,86
Drzewa wzmacniane	17,25	30,71
Sieci neuronowe	23,50	25,71

Zauważmy, że podczas tworzenia modeli pewne ustawienia dobierane są losowo (np. początkowe wagi sieci neuronowych). W związku z tym wyniki uzyskiwane przy wielokrotnym tworzeniu modeli mogą być nieco różne.

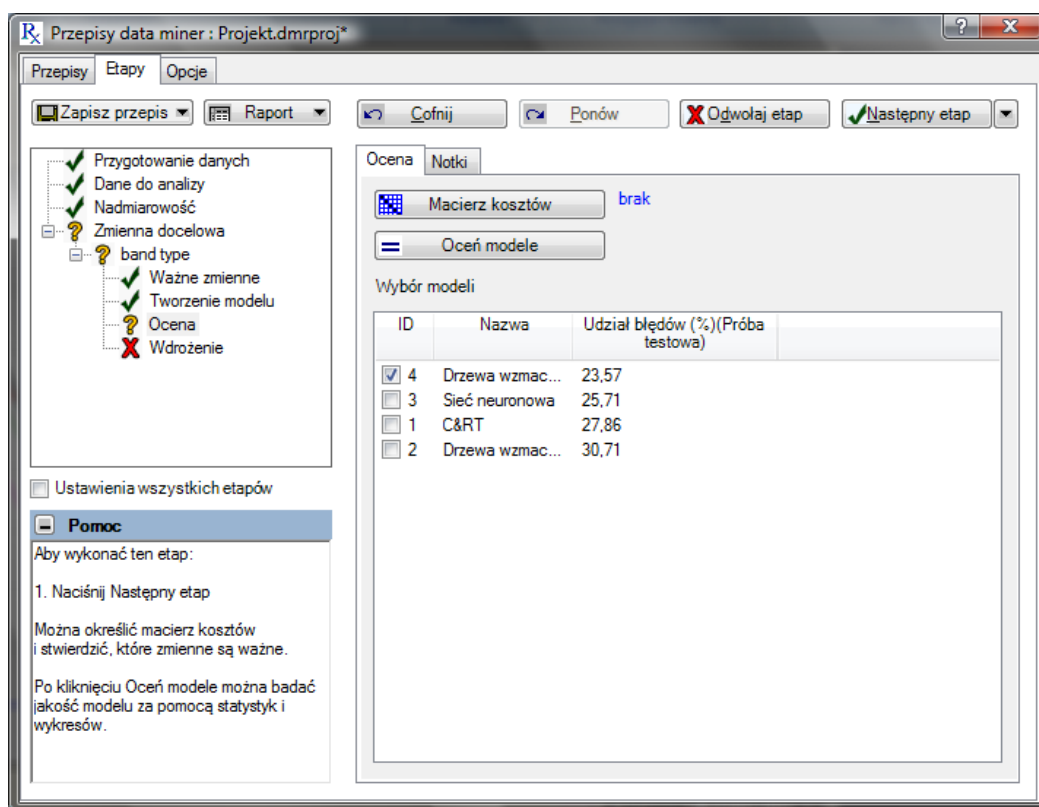
Drzewa wzmacniane na ogół dają bardzo dobre wyniki, a my dostaliśmy przeuczony model, najgorzej działający dla próby testowej. W związku z tym, spróbujemy poprawić ten model. Aby utworzyć dodatkowy model drzew wzmacnianych ze zmienionymi ustawieniami klikamy przycisk *Drzewa wzmacniane* i w oknie ustawień w polu *Maksymalna liczba węzłów* wpisujemy 3. Ponieważ chcemy ponownie dopasować tylko model drzew wzmacnianych, to odznaczamy pozostałe metody, po czym naciskamy przycisk *Twórz modele*.

Utworzony zostanie nowy model o numerze 4 (obok trzech zbudowanych wcześniej przy domyślnych ustawieniach). Nowy model ma stopę błędów w próbie uczącej równą 18,75% (tyle samo co sieci), a w próbie testowej 23,57%. Jeśli jako kryterium wyboru modelu przyjmujemy stopę błędnych przewidywań w próbie testowej to model drzew wzmacnianych jest najlepszy.



Rys. 10. Zmodyfikowane ustawienia dla drzew wzmacnianych.

Dokładniejszą ocenę działania modeli możemy wykonać na kolejnym etapie *Przepisu*. Przechodzimy do niego po naciśnięciu przycisku *Następny etap*. Następnie klikamy przycisk *Oceń modele*.



Rys. 11. Ocena modeli.

Aby przejrzeć statystyki i wykresy podsumowujące modele, naciskamy przycisk *Raport* i wybieramy polecenie *Wyniki wszystkich etapów*. Na ekranie pojawi się skoroszyt zawierający informacje o wykonanych do tej pory etapach *Przepisu* (zob. rys. 12).



1 Zmienna	2 Typ	3 Rola	4 Średnia	5 Odchylenie standardowe	6 Skośność
1 grain screened	Jakościowe	Wejście			
2 proof on ctd ink	Jakościowe	Wejście			
3 blade mfg	Jakościowe	Wejście			
4 paper type	Jakościowe	Wejście			
5 ink type	Jakościowe	Wejście			
6 direct steam	Jakościowe	Wejście			
7 solvent type	Jakościowe	Wejście			
8 type on cylinder	Jakościowe	Wejście			
9 press type	Jakościowe	Wejście			
10 press	Jakościowe	Wejście			
11 cylinder size	Jakościowe	Wejście			
12 paper mill location	Jakościowe	Wejście			
13 plating tank	Jakościowe	Wejście			
14 proof cut	Ilościowe	Wejście	45,0360825	8,57025085	0,36228645

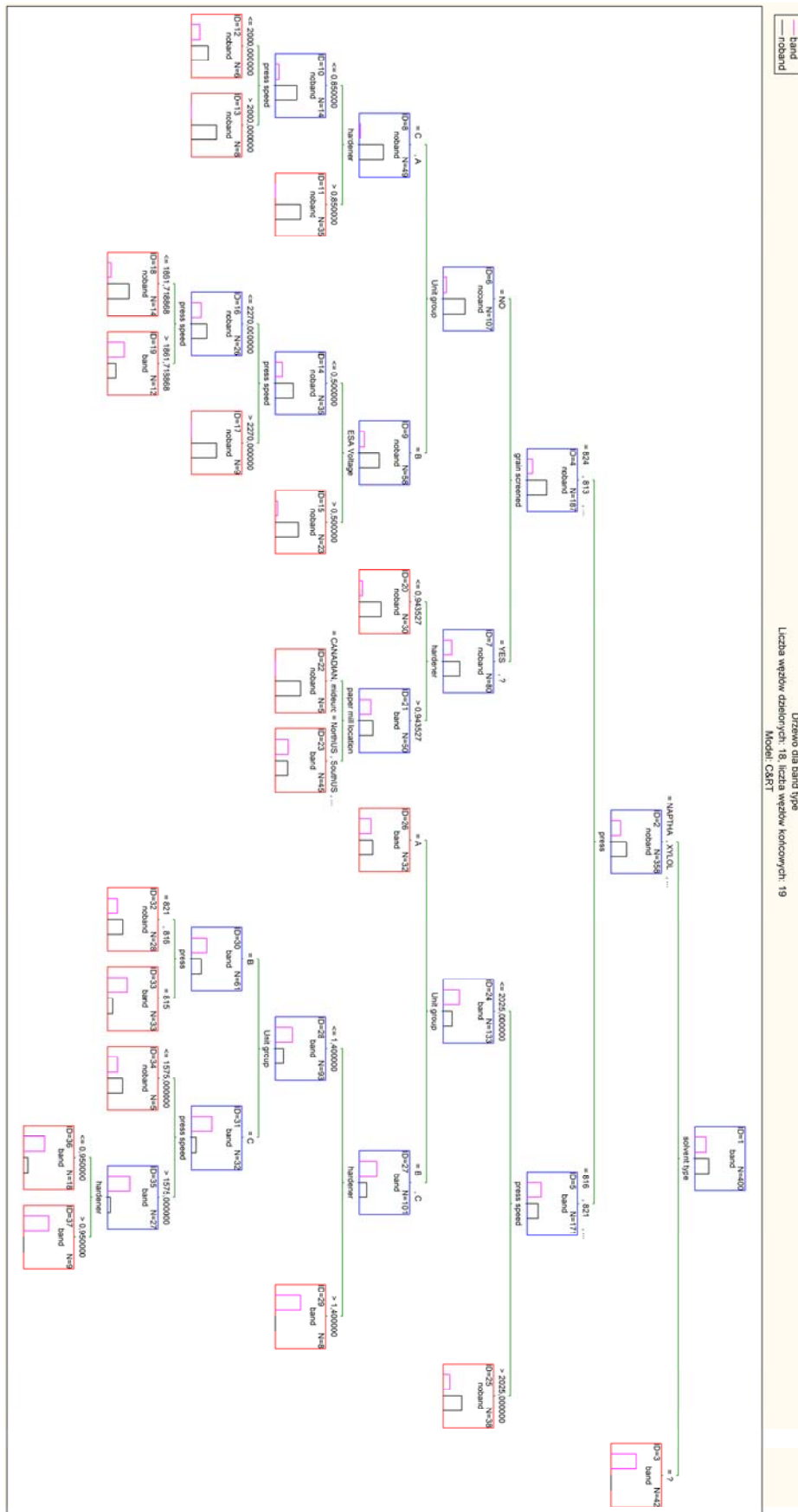
Rys. 12. Raport podsumowujący.

Zacznijmy od sprawdzenia, które zmienne zostały wybrane przez program do budowy modelu. Przechodzimy do węzła *Ważne zmienne* w folderze *Zmienna docelowa – band type*. Poniżej widzimy arkusz podsumowujący dobór zmiennych. W poniższej tabeli statystyka chi-kwadrat mówi jak silny jest wpływ konkretnej zmiennej, i tak: najmocniejszy jest wpływ zmiennej *ESA Voltage*. Druga kolumna w tabeli *Wartość p* informuje nas, jaka jest szansa, że zależność jest istotna: im wartość p jest mniejsza, tym większa jest pewność, że dana zmienna rzeczywiście wpływa na wyjście.

Najlepsze predyktory		
	Chi kwadrat	Wartość p
ESA Voltage	68,50188	0,000000
solvent type	58,62717	0,000000
proof on ctd ink	55,44276	0,000000
press	55,36386	0,000000
roller durometer	55,33793	0,000000
blade mfg	50,51595	0,000000
grain screened	48,95210	0,000000
paper mill location	45,65970	0,000000
press speed	38,46276	0,000014
hardener	31,80869	0,000215
paper type	31,41184	0,000000
Unit group	29,70422	0,000000

Rys. 13. Ważne zmienne.

Zobaczmy teraz, jak wygląda model dla drzewa C&RT. Drzewo znajduje się w folderze *Tworzenie modelu – 1 – C&RT* skoroszytu podsumowującego (przedstawia je rys. 14). Drzewo składa się łącznie z 37 węzłów, z czego 19 to węzły końcowe (zwane też liśćmi). Można powiedzieć, że cała zbiorowość została podzielona na 19 części, różniących się częstością występowania wad na wydrukach.



Rys. 14. Drzewo C&RT.

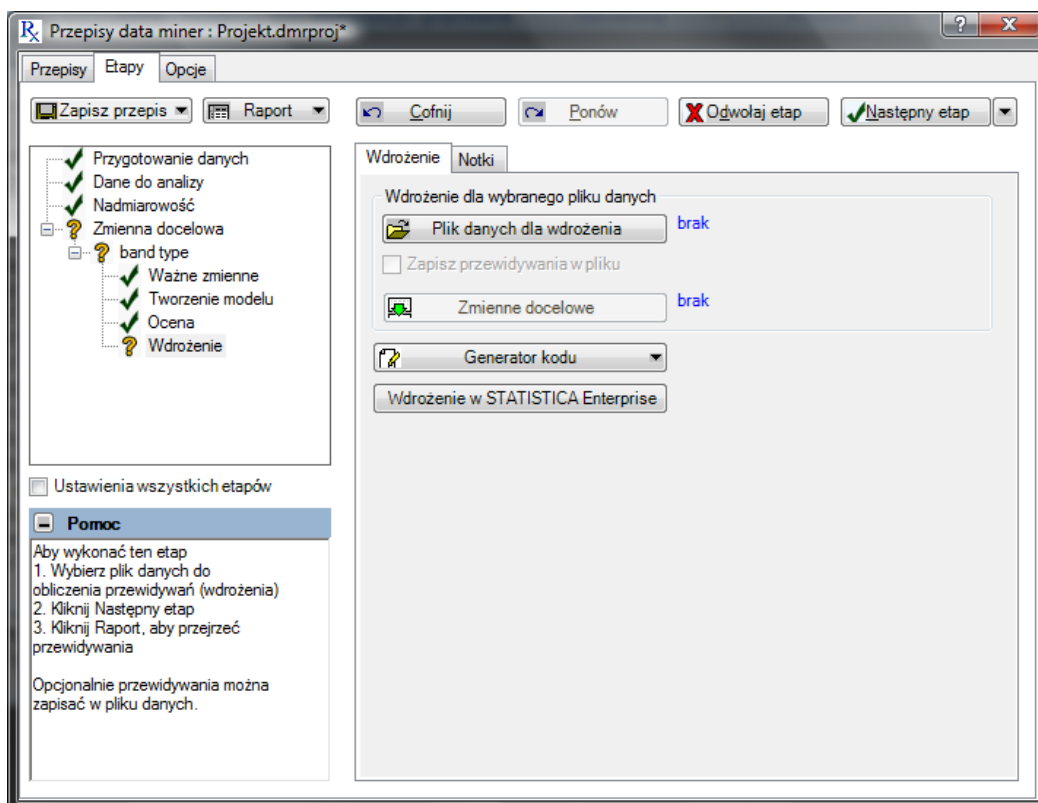


Warto zauważyć, iż model drzew klasyfikacyjnych jest bardzo łatwy do zrozumienia: po prostu przechodzimy w głąb drzewa i doczytujemy reguły prowadzące do poszczególnych liści (węzłów końcowych). I tak widzimy, że zawsze, gdy nieznanym był typ rozpuszczalnika (*solvent type* = ?), wydruki były wadliwe. *Banding* często występował również, gdy znany był typ rozpuszczalnika, wydruk wykonywano na prasach: 824, 813, 827, 828 lub 802, prędkość prasy była mniejsza od 2025, a wydruk wykonywano na jednostce z grupy A. W ten sposób możemy prześledzić wszystkie ścieżki prowadzące do wystąpienia pasów na wydruku i odkryć, co je powoduje. Dzięki temu drzewo może nam wskazać sekwencję zdarzeń prowadzącą do wystąpienia wady lub innych niekorzystnych zdarzeń.

Model możemy wykorzystać nie tylko do zidentyfikowania przyczyn występowania wad. Dwa inne częste zastosowania to:

- ◆ przewidywanie wyniku procesu,
- ◆ optymalizacja ustawień procesu.

Każde z powyższych zastosowań wymaga wdrożenia modelu, tzn. możliwości wyznaczenia jego przewidywań dla nowych danych. Po ocenie modeli kolejny etap *Przepisu* to właśnie wdrożenie modeli. Przed przejściem do tego etapu musimy podjąć decyzję, który model będziemy stosować.



Rys. 15. Stosowanie modelu.

Zarówno w przewidywaniu, jak i optymalizacji, zależy nam przede wszystkim na modelu najtrafniej przewidującym wynik procesu: w naszym przypadku, czy wystąpiły pasy na wydrukach, czy nie. W związku z tym do wdrożenia wybierzemy model nr 4 - drzewa



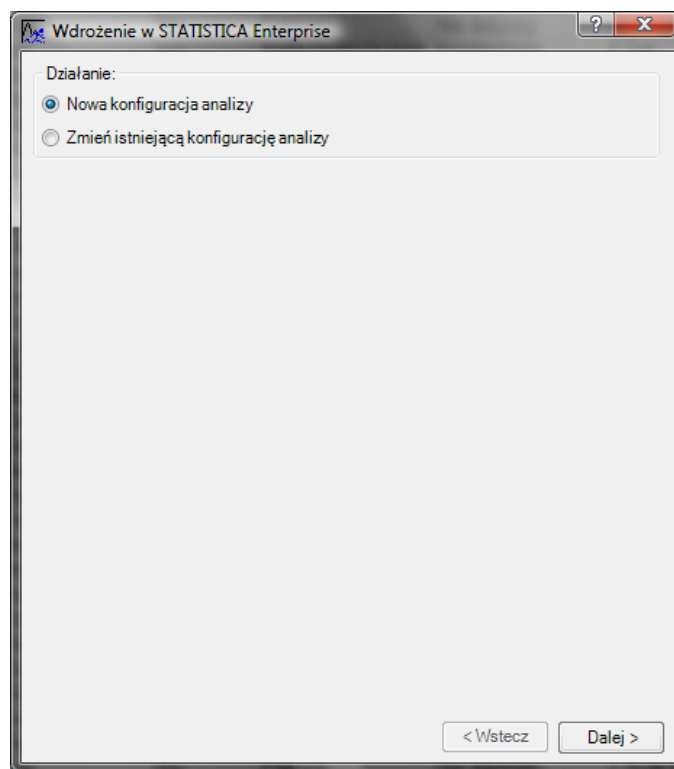
wzmacniane (jest on zresztą domyślnie wybrany do wdrożenia). Klikamy przycisk *Następny etap* i przechodzimy do *Wdrożenia*.

Na etapie *Wdrożenie* do dyspozycji mamy:

- ◆ zastosowanie modelu dla pliku danych,
- ◆ utworzenie kodu modelu w językach C i PMML,
- ◆ zastosowanie modelu w *STATISTICA Enterprise*.

My skorzystamy z ostatniej możliwości. Taki sposób wdrożenia jest korzystny i wygodny, ponieważ wszyscy uprawnieni użytkownicy systemu mają łatwo dostępne przewidywania modelu: aby je uzyskać, wystarczy po prostu wybrać odpowiedni szablon analizy i uruchomić go.

Naciskamy przycisk *Wdrożenie w STATISTICA Enterprise*. Na ekranie otworzy się okno, w którym decydujemy, czy tworzymy nową konfigurację, czy zmieniamy już istniejącą. Utworzymy nową konfigurację analizy.



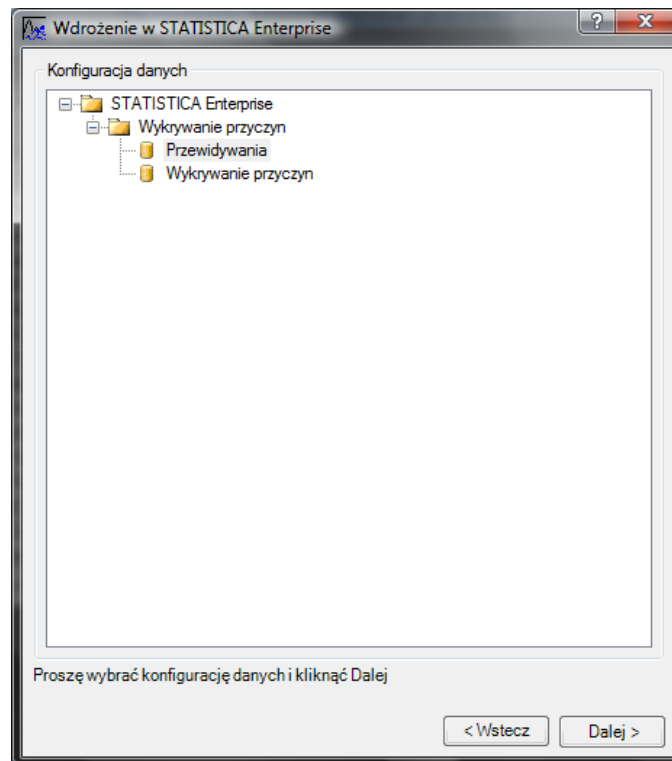
Rys. 16. Pierwszy etap wdrożenia.

W następnym etapie wskazujemy źródło danych, dla których stosowany będzie model, tzw. konfigurację danych. W naszym przypadku dla wdrożenia modeli przygotowano konfigurację danych *Przewidywania*.

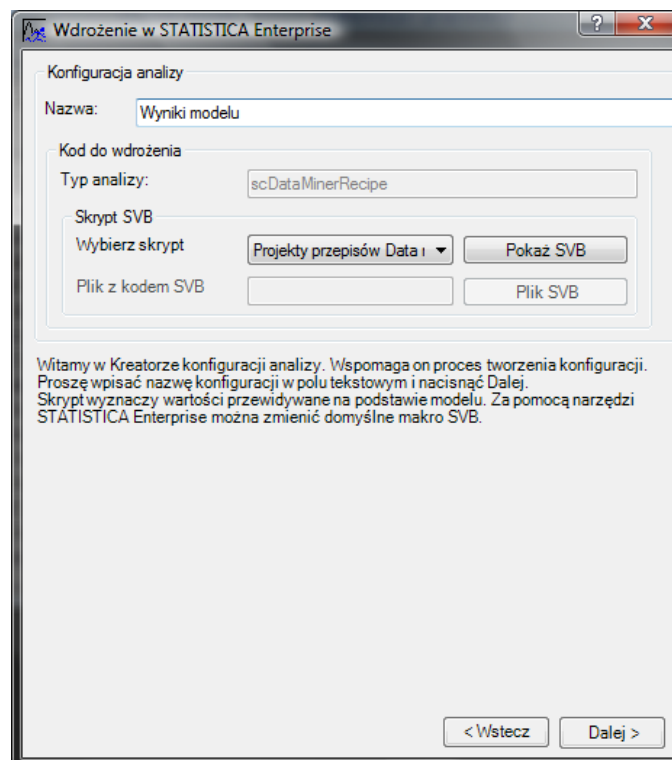
Następny krok to nazwanie konfiguracji wdrażającej model, nazwijmy ją *Wyniki modelu*.



Po podaniu nazwy konfiguracji określamy prawa dostępu do niej i na koniec jej lokalizację w systemie. Po wskazaniu foldera klikamy *Zakończ* – od tego momentu użytkownicy systemu mogą stosować model dla nowych danych „jednym kliknięciem”.



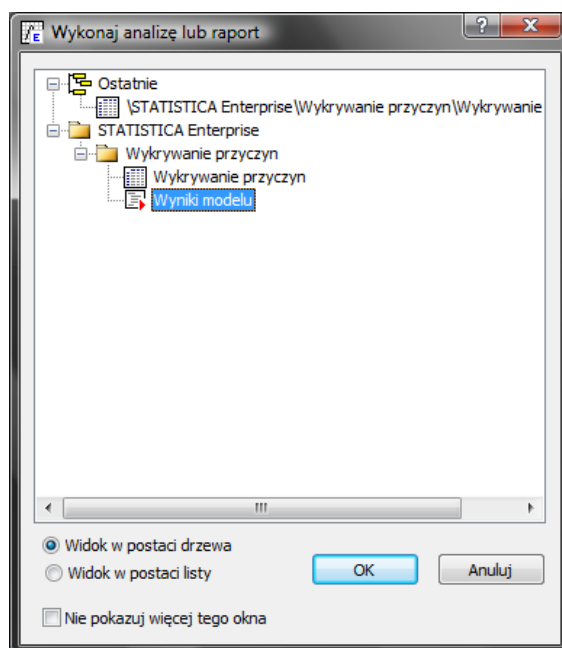
Rys. 17. Wybór konfiguracji danych.



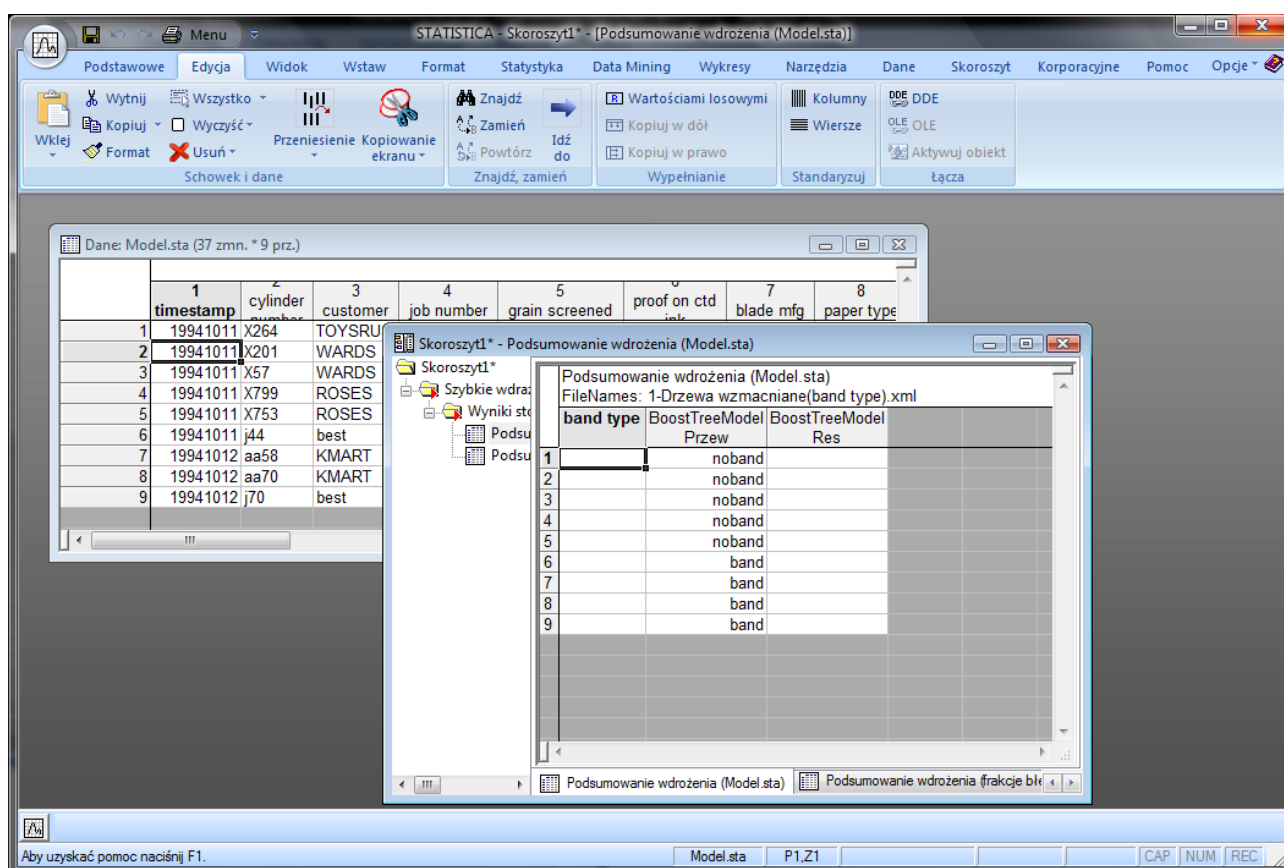
Rys. 18. Nazwanie konfiguracji analizy.



Dla przykładu zobaczmy, jak wygląda stosowanie modelu z punktu widzenia zwykłego użytkownika. Po uruchomieniu *STATISTICA* wybieramy polecenie *Korporacyjne – Uruchom analizę lub raport* i w oknie *Wykonaj analizę lub raport* wybieramy *Wyniki modelu*.



Rys. 19. Uruchamianie wdrożenia.



Rys. 20. Wdrożenie modelu w *STATISTICA*



Program automatycznie pobierze dane, wykona odpowiednie przekształcenia, zastosuje model, wyznaczy miary trafności i wyświetli wyniki w *STATISTICA*.

Ponieważ dane, dla których zastosowaliśmy model, nie zawierają informacji o wystąpieniu pasów na wydrukach (możemy przyjąć, że są to parametry nowych lub projektowanych procesów), to w wynikach wdrożenia nie mamy wartości rzeczywistej, ani trafności przewidywań, ale mamy to, na czym nam zależało: spodziewany wynik procesu dla nowych danych.

Podsumowanie

W niniejszym przykładzie przedstawiliśmy, jak korzystając z *Przepisów Data Miner* można stworzyć model procesu produkcyjnego, a następnie wykorzystać model do identyfikacji przyczyn wad i przewidywania wystąpienia wady nowych produktów. Użyliśmy danych dla przemysłu poligraficznego, ale podobny sposób postępowania można zastosować w niemal każdej gałęzi przemysłu, pod warunkiem, że dysponujemy danymi o procesach.

Przepisy Data Miner ułatwiają uzyskanie modelu, prowadząc użytkownika przez całą drogę od danych do wiedzy, wspierając nie tylko właściwe modelowanie, ale również przygotowanie danych do analizy, ocenę modelu i jego stosowanie. W środowisku *Przepisów* do stworzenia modeli użyliśmy drzew klasyfikacyjnych C&RT, sieci neuronowych i drzew wzmacnianych. Do identyfikacji przyczyn szczególnie użyteczne są drzewa klasyfikacyjne, ponieważ dają one zrozumiały model. Natomiast dwie pozostałe metody dają trafniejsze przewidywania.

Warto też zwrócić uwagę na ułatwienie pracy uzyskiwane dzięki *STATISTICA Enterprise*, zwłaszcza w sferze pobierania danych, przygotowania procedury stosowania modelu i uruchamiania go przez użytkowników.

Literatura

1. Hill T., Eames R., Lahoti S., *Finding Direction in Chaos*, Quality Digest, 12 (2009); <http://www.qualitydigest.com/magazine/2008/dec/article/finding-direction-chaos.html>
2. Berry M.J.A., Linoff G.S., *Mastering Data Mining*, Wiley Computer Publishing, 2000.
3. Asuncion A. & Newman D.J. (2007). UCI Machine Learning Repository [<http://www.ics.uci.edu/~mlearn/MLRepository.html>]. Irvine, CA: University of California, School of Information and Computer Science.
4. Demski T., *STATISTICA Enterprise jako platforma analityczna dla całej organizacji*, artykuł dostępny na stronie: <http://www.statsoft.pl/czytelnia/jakosc/wprowadzenie.html>.
5. Tadeusiewicz R., *Wprowadzenie do sieci neuronowych*, StatSoft Polska, 2001.
6. Koronacki J., Ćwik J., *Statystyczne systemy uczące się*, Wydawnictwo Naukowo-Techniczne, 2005.