POLITECHNIKA WARSZAWSKA

DYSCYPLINA NAUKOWA – INŻNIERIA CHEMICZNA
DZIEDZINA NAUK - NAUKI INŻYNIERYJNO-TECHNICZNE

Rozprawa doktorska

mgr inż. Dawid Szpadzik

Badanie możliwości wykorzystania narzędzi statystycznych w analizach jakościowych w obszarach produkcyjnych

Promotor
prof. dr hab. inż. Wioletta Raróg-Pilecka

WARSZAWA 2024

Podziękowania

Pragnę podziękować wszystkim, którzy przyczynili się do powstania tej pracy. Szczególne podziękowania kieruję do tych, którzy inspirowali mnie do dalszego rozwoju i poszukiwania nowych rozwiązań.

Wyrazy szczególnej wdzięczności kieruję do
Pani prof. dr hab. inż. Wioletty Raróg-Pileckiej
za bezprecedensowe, cierpliwe, humorystyczne
oraz wszechstronne wsparcie w realizacji niniejszej pracy doktorskiej
oraz umożliwienie mi rozwoju naukowego.

Administracji:
Pani inż. Ewie Szczygieł,
Pani mgr Annie Poskrobko
za pomoc w przejściu proceduralnych meandrów.

Najbliższym.

Codziennie dzieją się na świecie rzeczy, których nie da się wytłumaczyć na podstawie znanych nam praw. Codziennie, wywoławszy przedtem nieco szumu, rzeczy te są zapominane i ta sama tajemnica, która je przyniosła, zabiera je, a zagadka staje się zapomnieniem. Oto prawo: to co nie może być wyjaśnione, musi być zapomniane. Słoneczne światło nadal reguluje funkcjonowanie widzialnego świata. Obcość podpatruje nas z cienia.

Fernando Pessoa, *Księga niepokoju spisana przez Bernarda Soaresa, pomocnika księgowego w Lizbonie*,
przeł. Michał Lipszyc

Streszczenie

W niniejszej rozprawie przedstawiono tworzenie narzędzia statystycznego wykorzystywanego w analizach jakościowych, w obszarach produkcji chemicznej, z wykorzystaniem jedynie standardowego oprogramowania i komputera biurowego.

Opisano uczenie maszynowe oraz przykłady jego zastosowanie w obszarze technologii chemicznej łącznie z rysem historycznym. Przedstawiono powszechnie stosowane modele klasyfikacji binarnej, takie jak regresja logistyczna, losowy las decyzyjny, drzewa klasyfikacji i inne. Metodę wykorzystaną w niniejszych badaniach (algorytm naiwnego klasyfikatora Bayesa) opisano szerzej oraz zilustrowano ją przykładem. Przytoczono sposoby oceny skuteczności działania klasyfikatorów binarnych wykorzystujące tablicę pomyłek.

W związku z wdrożeniowym charakterem badań opisany został proces produkcji artykułów chemii gospodarczej. Polegał on na wytwarzaniu płynnych środków czystości na bazie roztworów wodnych w modelu szarżowym. Zwrócono uwagę na zarządzanie jakością w zakładzie wytwórczym (normy ISO), której celem było zapewnienie, że wyroby wprowadzane na rynek są zgodne z deklaracjami producenta oraz obowiązującymi normami prawnymi. Uwzględniony został aspekt finansowy (kosztochłonności) utrzymania laboratorium fizykochemicznego, który składał się z wydatków na materiały zużywalne (np. odczynniki, pipety, rękawiczki) oraz zatrudnienie wykwalifikowanego personelu. Praca badawcza została podjęta w ramach działań ciągłego doskonalenia procesu produkcyjnego oraz optymalizacji kosztów wytwarzania w przemyśle chemicznym. Dążyły one ku zapewnieniu odpowiedniej marżowości produktu oraz konkurencyjnej ceny dla konsumenta. Działania te były szczególnie istotne w kontekście presji inflacyjnej wywołanej pandemią COVID-19 oraz sytuacją geopolityczną w Europie.

W części eksperymentalnej, w pierwszej fazie przeanalizowano proces wytwórczy oraz technologiczny realizowany u współpracującego przedsiębiorcy. Ustalone zostały wymagania stawiane opracowywanemu narzędziu oraz kryteria sukcesu. Wyszczególnione zostało kryterium badawcze (narzędzie klasyfikujące osiąga dokładność lepszą niż losowa) oraz kryterium wdrożeniowe (dokładność przypisania klasy na poziomie minimum 95%). Opisano działanie stworzonej aplikacji oraz wbudowane funkcje, których celem było wypełnienie wymagań postawionych w poprzedzającym etapie. Klasyfikator wyposażono w szereg parametrów, które mogą być modyfikowane przez użytkownika.

Drugim etapem części eksperymentalnej była optymalizacja działania algorytmu uczenia maszynowego (tzw. trenowanie). W pierwszej kolejności został ustalony ogólny punkt odniesienia – ustawienia referencyjne. Następnie każdy parametr algorytmu był iteracyjnie modyfikowany w celu zbadania charakterystyki jego wpływu na jakość klasyfikacji. W kolejnym kroku wysterowano wszystkie argumenty jednocześnie, przy użyciu tych nastaw, dla których uprzednio uzyskano najlepsze wartości dokładności oraz współczynnika korelacji. Ostatnim krokiem było zbadanie możliwości ograniczenia liczby analiz fizykochemicznych, w których zużywane są materiały jednorazowe oraz odczynniki.

Udowodniono, że jest możliwe opracowanie narzędzia wykorzystującego model uczenia maszynowego, pracującego na komputerze osobistym, stworzonego przy użyciu oprogramowania biurowego Microsoft Office. Potwierdzono możliwości wykorzystania narzędzi statystycznych w kontroli jakości w przemyśle chemicznym (cel badawczy pracy doktorskiej). Dokładność opracowanego narzędzia klasyfikującego była większa od 50% – tym samym lepsza od przypisania losowego. Został osiągnięty wymagany do wdrożenia poziom dokładności (minimum 95%), przy jednoczesnym zachowaniu kryterium przyznania ocen próbkom ze zbioru treningowego (minimum 80%). W populacji testów osiągnięta najmniejsza dokładność wyniosła 97,85%. Prace optymalizacyjne wykazały, że liczba kosztochłonnych badań fizykochemicznych może zostać zredukowana, co dodatkowo może wpłynąć na zwiększenie bezpieczeństwa personelu laboratoryjnego.

W toku prac badawczych wykazano, że korekcje matematyczne nie przyniosły pożądanych efektów. Niwelowanie wpływu braku wystąpień obserwacji (wygładzenia Laplace'a) spowodowało znaczny pogorszenie wskaźników jakości klasyfikacji. Nie zaobserwowano również znaczących zmian w metrykach przy zastosowaniu dyskretyzacji zmiennych rzeczywisty w oparciu o odchylenie standardowe.

W przedstawionej rozprawie doktorskiej udowodniono, że jest możliwe zbudowanie algorytmu uczenia maszynowego z wykorzystaniem standardowego oprogramowania (Microsoft Office) oraz komputera biurowego. Aplikacja spełniła wymogi, aby uzyskać potencjał wdrożeniowy w zakładzie wytwórczym przemysłu chemicznego Reckitt Benckiser Production (Poland) sp. z o.o. Dokładność klasyfikacji osiągnęła poziom 99,85%, a współczynnik korelacji 0,9758. Omówiono parametry sterujące algorytmem oraz ich wpływ na skuteczność oznaczenia klasy. Reasumując, uwidoczniono sposobność zastosowania uczenia maszynowego w technologii produkcji płynnych środków czystości oraz udowodniono, że jego opracowanie nie wymaga specjalistycznych narzędzi komputerowych. Możliwe jest jego wykorzystanie w powszechnych procesach produkcyjnych środków chemii

gospodarczej, zmniejszając koszty związane z utrzymaniem laboratorium kontroli jakości. Tym samym optymalizacja działań inżynierów chemików rozszerza się o modelowanie matematyczne całego zakresu wykonywanych zadań w przemyśle chemicznym, nie zaś jedynie na pojedynczych metodach fizykochemicznych.

Słowa kluczowe: uczenie maszynowe, naiwny klasyfikator Bayesa, klasyfikacja binarna, przemysł chemiczny, kontrola jakości w przemyśle chemicznym, technologia produkcji środków chemii gospodarczej

Abstract

This dissertation presents the creation of statistical tools in quality control analysis, in household chemicals manufacturing industry, using standard software and an office computer.

Machine learning and examples of its application in the field of chemical engineering are described including a historical overview. Commonly used binary classification models such as logistic regression, random decision forest, classification trees and others are presented. The method used in this study (naive Bayes classifier algorithm) is described in more detail and illustrated with an example. Ways of evaluating the performance of binary classifiers using a confusion table are cited.

Due to the implementation nature of the research, the process of manufacturing housekeeping chemicals was described. It consisted of manufacturing liquid cleaning products based on aqueous solutions in a batch model. Attention was paid to quality management at the manufacturing plant (ISO standards), the purpose of which was to ensure that the products placed on the market comply with the manufacturer's declarations and applicable legal standards. Consideration was given to the financial aspect (cost intensity) of maintaining a physicochemical laboratory, which consisted of expenditures for consumables (e.g. reagents, pipettes, gloves) and with the employment of qualified personnel. The research work was undertaken as part of efforts to continuously improve the production process of household chemicals and optimize manufacturing costs in chemical industry. They sought to ensure adequate product margins and a competitive price for the consumer. These activities were particularly important in the context of inflationary pressures caused by the COVID-19 pandemic and the geopolitical situation in Europe.

In the experimental part, the first phase analyzed the manufacturing process carried out at the cooperating entrepreneur. The requirements for the tool under development and the success criteria were established. A research criterion (the classifier tool achieves accuracy better than random) and an implementation criterion (class assignment accuracy of at least 95%) were specified. The operation of the created application was described, as well as its functions aimed at fulfilling the requirements set in the preceding stage. The classifier is equipped with a number of parameters that can be modified by the user.

The second stage of the experimental part was to optimize the performance of the machine learning algorithm (known as training). First, a general benchmark - reference settings - was established. Then each parameter of the algorithm was iteratively modified to study the characteristics of its effect on the quality of classification. In the next step, all arguments were set off simultaneously, using those settings for which the best values of accuracy and correlation coefficient were previously obtained. The final step was to investigate the possibility of reducing the number of physicochemical analyses that use disposable materials and reagents.

It was proved that it is possible to develop a tool using a machine learning model, working on a personal computer, created using Microsoft Office software. It was confirmed that it is possible to use statistical tools in quality control analysis, in production areas (the research goal of the doctoral thesis). The accuracy of the developed classifier tool was greater than 50% - thus better than random assignment. The level of accuracy required for implementation was achieved (a minimum of 95%), while maintaining the criterion for assigning classes to samples from the training set (a minimum of 80%). In the test population, the lowest accuracy achieved was 97.85%. The optimization work showed that the number of cost-intensive physical-chemical tests can be reduced, which can further improve the safety of laboratory personnel.

The research work showed that mathematical corrections did not have the desired effect. Leveling the effect of the absence of occurrences of observations (Laplace smoothing) resulted in a significant deterioration of classification quality indicators. Also, no significant changes in metrics were observed when using standard deviation-based discretization of real variables.

The presented dissertation proved that it is possible to build a machine learning algorithm using standard software (Microsoft Office) and an office computer. The application met the requirements to achieve implementation potential in a manufacturing plant of Reckitt Benckiser Production (Poland). Classification accuracy reached 99.85%, and the correlation coefficient reached 0.9758. The parameters controlling the algorithm and their impact on the effectiveness of class determination were discussed. In summary, the capabilities of machine learning tools were learned and proved that their application does not require specialized computer tools and can be used in standard manufacturing processes of liquids household products, reducing the costs associated with maintaining a quality control laboratory. Thus, the optimization of the activities of chemical engineers expands to incorporate mathematical modeling of the entire range of work to be performed, rather than merely on isolated physicochemical methods.

Keywords: machine learning, naive Bayes classifier, binary classification, chemical industry, quality control in chemical industry, production technology of household chemicals

Spis treści

1. Układ pracy

Przedstawiona rozprawa doktorska zawiera sześć przedmiotowych rozdziałów: Wstęp, Przegląd literatury, Część eksperymentalną, Wyniki badań, Dyskusję wyników oraz Wnioski.

Wstęp (Rozdział 3) stanowi tło i krótkie wprowadzenie w tematykę niniejszej rozprawy. Przedstawiono trendy globalizacji, zwiększenia kompleksowości oraz cyfryzacji, które obserwuje się w przemyśle. Na zakończenie tego rozdziału sformułowano cel pracy.

W Przeglądzie literatury (Rozdział 4) przedstawiono aktualny stan wiedzy w obszarach podjętych w rozprawie. Omówiono dziedzinę uczenia maszynowego, rys historyczny oraz zastosowanie tejże dziedziny w obszarze technologii chemicznej. Przedstawiono modele klasyfikacji binarnej wraz z metodami oceny ich jakości. W związku z wdrożeniowym charakterem pracy badawczej, opisano typ procesu produkcyjnego środków chemii gospodarczej, który był realizowany w zakładzie wytwórczym Reckitt Benckiser Production (Poland), a którego praca dotyczy.

W Części eksperymentalnej przedstawiono wyłanianie koncepcji narzędzia statystycznego wraz z wymaganiami stawianymi przedmiotowi wdrożenia (Rozdział 5.2). W Rozdziale 5.3 pokazano interfejs graficzny oraz opisano funkcje opracowanej aplikacji. Omówiono przyjęty model testowania klasyfikacji (Rozdział 5.4), tzw. trenowania algorytmu.

Rozdział 6 Wyniki badań, zawiera rezultaty przeprowadzonych badań. Opisano realizację modelu testowania. Omówiono wyniki, czyli zależności pomiędzy modyfikacją parametrów sterujących algorytmem, a jakością klasyfikacji. Przedstawiono optymalne wartości nastaw. Zaproponowano redukcję liczby analiz fizykochemicznych wykonywanych w ramach kontroli wytwarzanych produktów.

Rozdział 7 poświęcono dyskusji otrzymanych wyników.

W ostatnim rozdziale przedmiotowym (Rozdział 8), przedstawiono wnioski, jakie sformułowano w oparciu o wyniki przeprowadzonych prac badawczych.

2. Spis skrótów

ABS – kwas alkilobenzenosulfonowy

ACC – (ang. *accuracy*) dokładność

ACC_{MAX} – maksymalna wartość dokładności (ACC) uzyskana w analizowanej grupie testów

ACC_{MIN} – minimalna wartość dokładności (ACC) uzyskana w analizowanej grupie testów

AFM – (ang. *Atomic Force Microscopy*) mikroskopia sił elektronowych

AN – (ang. *Actual Negative*) liczba próbek w zbiorze treningowym z etykietę negatywną

AP – (ang. *Actual Positive*) liczba próbek w zbiorze treningowym z etykietę pozytywną

ASR – wskaźnik jaka część próbek z dostępnych do klasyfikacji w zbiorze treningowym otrzymała klasę wyrażany w procentach

ASR_{MAX} – maksymalna wartość wskaźnika ASR uzyskana w analizowanej grupie testów

ASR_{MIN} – minimalna wartość wskaźnika ASR uzyskana w analizowanej grupie testów

DZPR – wskaźnik jakości produkcji: dobre za pierwszym razem

EM – (ang. *Electron Microscopy*) mikroskopia elektronowa

EOG – Europejski Obszar Gospodarczy

EWD – (ang. *Equal Width Discretization*) dyskretyzacja na równe przedziały

F1-score – średnia harmoniczna precyzji i czułości

FMCG – (ang. *Fast Moving Consumer Goods*) wyroby szybkozbywalne

FMEA – (ang. *Failure Mode Effective Analysis*) analiza ryzyk, przyczyn i skutków

FN – (ang. *False Negative*) liczba klasyfikacji fałszywie negatywnych

FNR – (ang. *False Negative Rate*) częstość fałszywych odkryć

FP – (ang. *False Positive*) liczba klasyfikacji fałszywie pozytywnych

FPR – (ang. *False Positive Rate*) częstość fałszywych alarmów

FPR_{MAX} – maksymalna wartość częstości fałszywych alarmów (FPR) uzyskana w analizowanej grupie testów

FPR_{MIN} – minimalna wartość częstości fałszywych alarmów (FPR) uzyskana w analizowanej grupie testów

ISO – (ang. *International Standards Organization*) międzynarodowa organizacja normalizacyjna

kNN – (ang. *k-Nearest Neighbors*) algorytm *k*-najbliższych sąsiadów

MCC – (ang. *Matthews Correlation Coefficient*) współczynnik korelacji Matthewsa

MCC_{MAX} – maksymalna wartość współczynnika korelacji Matthewsa (MCC) uzyskana w analizowanej grupie testów

MCC_{MIN} – minimalna wartość współczynnika korelacji Matthewsa (MCC) uzyskana w analizowanej grupie testów

MS – (ang. *Mass Spectrometry*) spektroskopia mas

NMR – (ang. *Nuclear Magnetic Resonance*) magnetyczny rezonans jądrowy

OTC – (ang. *Over The Counter*) leki sprzedawane bez recepty

PDCA – (ang. *Plan, Do, Check, Act*) cykl ciągłego doskonalenia, cykl Deminga

PN – (ang. *Predicted Negative*) liczba próbek zaklasyfikowanych jako negatywne

PP – (ang. *Predicted Positive*) liczba próbek zaklasyfikowanych jako pozytywne

PPV – (ang. *Positive Predicted Value*) precyzja pozytywna

ROC – (ang. *Receiver Operating Characteristic*) graficzna reprezentacja efektywności modelu predykcyjnego poprzez wykreślenie charakterystyki jakościowej klasyfikatorów binarnych

SLES – anionowy związek powierzchniowo czynny (Sodium laureth sulfate)

SZJ – System Zarządzania Jakością

TN – (ang. *True Negative*) liczba klasyfikacji prawdziwie negatywnych

TNR – (ang. *True Negative Rate*) specyficzność

TP – (ang. *True Positive*) liczba klasyfikacji prawdziwie pozytywnych

TPR – (ang. *True Positive Rate*) czułość

TPR_{MAX} – maksymalna wartość czułości (TPR) uzyskana w analizowanej grupie testów

TPR_{MIN} – minimalna wartość czułości (TPR) uzyskana w analizowanej grupie testów

VBA – (ang. *Visual Basic for Applications*) język programowania wbudowany w aplikacje pakietu Microsoft Office

VUCA – (ang. *Volatility, Uncertainty, Complexity and Ambiguity*) określenia stosowane do opisywania gospodarki: zmienność, niepewność, złożoność, niejednoznaczność

3. Wstęp

Obecnie zachodząca transformacja technologiczna i organizacyjna przedsiębiorstw określana jest jako czwarta rewolucja przemysłowa lub Przemysł 4.0 (ang. *Industry 4.0*). Od lat 70. XX wieku zachodzi zmiana organizacji produkcji w gospodarce. Obserwuje się, że firmy przekształcają się w wielonarodowe koncerny, połączone skomplikowanym łańcuchem dostaw (np. BASF, Merck, Dupont, 3M). Intuicyjnie postrzega się, że wraz z procesem globalizacji podąża standaryzacja. W rzeczywistości, chociaż globalizacja sprawiła, że niemal na całym świecie dostępne są jednorodne towary, różnice między fabrykami (procesami produkcyjnymi) nie tylko utrzymywały się, ale także pogłębiły. Konkurencyjność (wynikająca z globalnego rynku) oraz duże skomplikowanie procesów (prowadzenie operacji w wielu państwach) sprawiło, że firmy poszukują rozwiązań w narzędziach komputerowych. Aspektami, które przemawiają za cyfryzacją są: niskie koszty, szybkie efekty, przejrzystość działania oraz niezawodność [1–4].

Obserwowany jest wzrost liczby oferowanych artykułów wytwarzanych przez firmy produkcyjnie. Ma on swoje źródło w zwiększonej konkurencji, globalizacji, rosnącej liczbie klientów oraz popycie na unikalne wyroby (np. możliwość personalizacji) [5,6]. Przykładem tego zjawiska jest duński producent klocków dla dzieci – LEGO. W latach 1997–2004 firma podwoiła liczbę unikalnych klocków do ponad 12 000. Wkroczyła ona także na nowe obszary, takie jak gry komputerowe, odzież dziecięca i parki tematyczne. Wraz ze wzrostem różnorodności produktów zwiększyło się skomplikowanie procesów operacyjnych LEGO [6]. Analogiczny trend obserwowany jest także w kategoriach elektroniki użytkowej, środków czystości, farmaceutyków czy żywności [6–8].

Nierozłącznym elementem procesów wytwórczych w przemyśle chemicznym jest kontrola jakości. Ma ona miejsce na każdym etapie produkcji. Kontrolowane są materiały dostarczane przez dostawców, półprodukty oraz wyroby gotowe. Obowiązek zapewnienia, że dany wyrób spełnia zakresy prezentowanej specyfikacji pochodzi z trzech źródeł: wymogów prawnych, oczekiwań klientów oraz wewnętrznych regulacji producenta. Wraz ze wzrostem liczby wytwarzanych artykułów rośnie liczba wymaganych kontroli i analiz laboratoryjnych.

Tak więc obserwowane jest zwiększone zapotrzebowanie na wysoko wykwalifikowany personel. Rosną także koszty realizacji kontroli jakości w fabrykach (np. zwiększenie zużycia odczynników laboratoryjnych) [8–12].

Transformacja przemysłu, w ramach dążenia producentów ku Przemysłowi 4.0, skupia się na robotyzacji parku maszynowego oraz cyfryzacji procesów produkcyjnych. Istotne są również działania takie jak akwizycja i analiza danych, które służą do podejmowania decyzji biznesowych na dowodach empirycznych [1,8,13]. Szeroka dostępność komputerów oraz gwałtowny rozwój nauk informacyjnych (np. uczenie maszynowe, sztuczna inteligencja) otwiera nowe możliwości cyfryzacji elementów procesów produkcyjnych. Komputery osobiste (laptopy), które często są podstawowym wyposażenia stanowiska specjalisty, posiadają odpowiednie parametry i narzędzia do analizy danych i automatyzacji procesów. Decyzje, które wcześniej podejmował człowiek na bazie wykształcenia oraz wieloletniego doświadczenia mogą zostać sprowadzone do algorytmów ucznia maszynowego. Tak więc drogą do optymalizacji procesów produkcyjnych w fabrykach jest cyfryzacja i automatyzacja, nawet w tak kluczowej dziedzinie jak kontrola jakości [9,14–18].

Nieodłączną rolą inżynierów chemików jest projektowanie stanowisk laboratoryjnych, opracowywanie metod oraz zakresów eksperymentów. Standaryzacja oraz optymalizacja działań w laboratorium fizykochemicznym wymaga dogłębnego zrozumienia przedmiotu badań oraz celu ich wykonywania. W nowoczesnych laboratoriach automatyzacja ma szerokie zastosowanie, wykorzystywane są np. podajniki próbek, aparaty miareczkujące, reaktory badawcze. Projektowanie i tworzenie nowoczesnego stanowiska chemicznego w oparciu o najnowszą wiedzę prowadzi do integracji chemii, automatyki oraz informatyki. Obserwowany jest bezprecedensowy wzrost zastosowania modeli matematycznych (uczenie maszynowe, sztuczna inteligencja) w tradycyjnych dyscyplinach naukowych i inżynieryjnych. Większa dostępność oprogramowania i narzędzi sprzętowych do wdrażania sztucznej inteligencji zmniejszyła bariery w stosowaniu jej w badaniach chemicznych. Ponadto wielu badaczy nauczyło się technik generowania i obsługi danych, które następnie analizują z wykorzystaniem nowoczesnych algorytmów uczenia maszynowego [19–22].

3.1. Cel pracy

Uczenie maszynowe oraz sztuczna inteligencja są narzędziami, które zyskują coraz większe uznanie i zastosowanie w wielu dziedzinach nauki i życia codziennego. Rozwiązania te wymagają nowoczesnego oprogramowania komputerowego, zakupu licencji oraz odpowiednich zasobów obliczeniowych. Powstaje więc przeszkoda, która często uniemożliwia zastosowanie technologii z kategorii uczenia maszynowego w projektach rozwojowych np. w zakładach przemysłu chemicznego. Z jednej strony teoria Bayesa cieszy się powszechnym zastosowaniem w narzędziach np. do odfiltrowywania spamu czy szacowania zdolności kredytowej. Z drugiej strony istnieje jednak zauważalny brak tego typu narzędzi w procesach oceny jakości (fizykochemicznej) produktów w trakcie procesu produkcyjnego. Najważniejszą tezą niniejszej pracy jest stwierdzenie, że możliwe jest opracowanie narzędzia pracującego na komputerze osobistym, stworzonego przy użyciu oprogramowania biurowego Microsoft Office oraz wykorzystującego algorytm uczenia maszynowego – naiwny klasyfikator Baysa, w celu predykcji wyników klasyfikacji próbek produktu w ramach kontroli jakościowej w przemyśle chemicznym (wytwarzanie płynnych artykułów chemii gospodarczej). Cel pracy to opracowanie takiego narzędzia i udowodnienie postawionej tezy.

4. Przegląd literatury

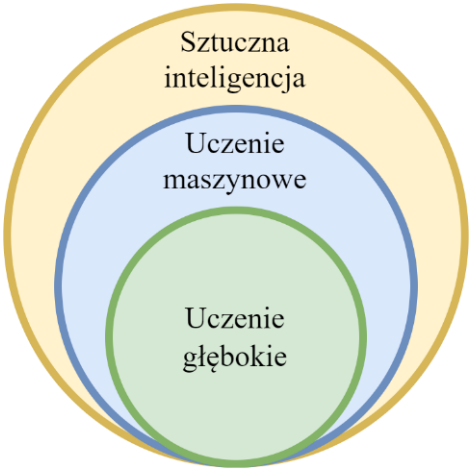
W tej części rozprawy dokonano przeglądu literaturowego obecnego stanu wiedzy na temat uczenia maszynowego z uwzględnieniem klasyfikatorów binarnych. Przedstawiono zarys historyczny wykorzystania modeli predykcyjnych w technologii chemicznej, który uzupełniono przykładami zastosowań. W nawiązaniu do wdrożeniowego charakteru tematyki badawczej opisany został model produkcji seryjnej (ang. *batch manufacturing*) w przemyśle chemicznym oraz istotę zarządzania jakością w procesie wytwórczym artykułów chemii gospodarczej. Zreferowane zostały podstawowe algorytmy klasyfikacji binarnej oraz technologie używane do jej wdrożenia. W podsumowaniu uwypuklone zostały główne wnioski wynikające z badań literaturowych.

4.1. Definicja uczenia maszynowego

W nauce istnieją pojęcia, które są intuicyjnie rozumiane oraz szeroko stosowane w praktyce, pomimo że nie posiadają skończonej i wyczerpującej definicji. Przykładem takiego przypadku jest *wirus*, odnośnie do którego ciągle nie istnieje konsensus naukowy, czy owe indywiduum biologiczne należy do świata organizmów żyjących, czy też nie [23]. Pojęcie *uczenia maszynowego* również zalicza się do takiej kategorii pojęć, ponieważ istnieje wiele teorii oraz towarzyszących im definicji, których celem jest zdefiniowanie, co zawiera się w zakresie procesów i rozwiązań, które kryją się pod parasolem pojęciowym uczenia maszynowego (ang. *machine learning*) [24–26].

Sztuczna inteligencja (ang. *artificial intelligence*) jest obszarem badań w dziedzinach nauk podstawowych i wdrożeniowych, których celem jest stworzenie maszyny imitującej inteligencję człowieka [26,27]. W owym obszarze wydzielony został podobszar, który zajmuje się wykreowaniem u maszyn zdolności do rozwiązywania szczegółowego problemu – uczenie maszynowe. Kolejnym wyodrębnionym podzakresem jest uczenie głębokie (ang. *deep learning*) [28]. Najwyższy obszar skupia w sobie całość działań, które opracowują odpowiednik inteligencji ludzkiej bądź jej fragmenty. Dwa niższe poziomy skupiają prace

mające na celu stworzenie narzędzi rozwiązujących zdefiniowany problem (np. wygranie z człowiekiem w warcaby) [24,29]. Uczenie głębokie zostało wydzielone, ponieważ w przeciwieństwie do konwencjonalnego (nadzorowanego) uczenia maszynowego w danych, które służą do trenowania algorytmów nie muszą być zdefiniowane oraz oznaczone wzorce. Algorytmy te są tworzone w taki sposób by same wykrywały schematy w zbiorze danych pierwotnych [30]. Na Rysunku 4.1 przedstawiono koncepcję współzależności pomiędzy sztuczną inteligencją, uczeniem maszynowym oraz uczeniem głębokim [31].

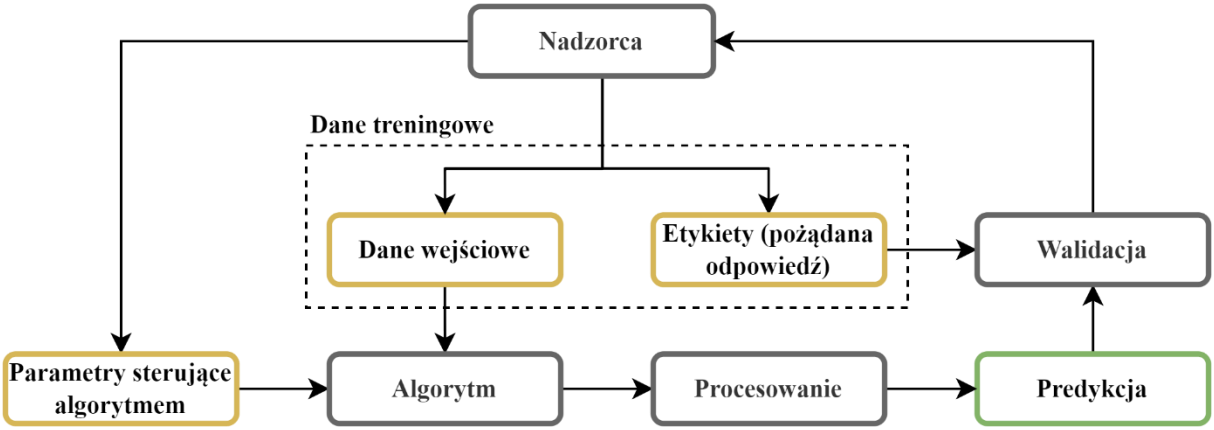


Rysunek 4.1. Diagram ilustrujący obszary sztucznej inteligencji [31].

W najnowszej literaturze ukucie pojęcia uczenia maszynowego oraz powszechnie stosowanej definicji („zdolność komputerów do uczenia się bez programowania nowych umiejętności wprost") przypisuje się Arthurowi Samuelowi z firmy IBM, który w 1959 roku udowodnił, że program komputerowy może wygrać z człowiekiem w warcaby [32].

W klasycznym rozumieniu uczenie maszynowe są to narzędzia (szczególnie oprogramowanie), których zadaniem jest rozwiązanie zdefiniowanego problemu. Są one tworzone w taki sposób, aby mogły się dostrajać do danych, które przetwarzają. Jednak, niezbędny jest proces uczenia (trenowania), aby algorytm uzyskał zdolności predykcyjne. Polega on na iteracyjnej analizie danych, którym towarzyszą przypisane, pożądane odpowiedzi algorytmu tzw. etykiety. Dzięki takiej strukturze danych na etapie trenowania możliwe jest porównanie wyników obliczeń modelu z rzeczywistą wartością. Jeżeli wskaźniki oceniające własności predykcyjne algorytmu są niezadowalające to następuje korekcja parametrów sterujących. Proces ten jest powtarzany, aż do uzyskania pożądanych wartości wskaźników [27,33–35]. W związku z tym, że dane wejściowe są inicjalnie przygotowywane poprzez przypisywanie im odpowiednich etykiet, ten rodzaj algorytmów nazywany jest również

uczeniem maszynowym nadzorowanym, gdzie opracowanie danych najczęściej wykonywane jest przez człowieka [36]. Rysunek 4.2. przedstawia proces trenowania algorytmu, w którym nadzorca przypisuje etykiety do danych wejściowych oraz ustawia odpowiednie parametry sterujące algorytmem. Walidacja jest procesem porównania etykiet przypisanych przez człowieka z predykcją algorytmu, na podstawie którego nadzorca koryguje parametry sterujące.



Rysunek 4.2. Schemat blokowy procesu trenowania algorytmu w modelu nadzorowanego uczenia maszynowego [36].

W związku z faktem, że skuteczność procesu uczenia jest proporcjonalna do ilości danych w zbiorze treningowym, to nadzorowane uczenie maszynowe jest łatwo aplikowalne w zagadnieniach badawczych, które posiadają już odpowiednie bazy danych z etykietami [37,38]. Przykładami takich zbiorów danych są: bazy dokumentów tekstowych z przypisanymi im klasami; bazy pacjentów z objawami oraz przypisanymi im jednostkami chorobowymi. W pierwszym przykładzie algorytm jest trenowany na podstawie analizy słów w dokumencie oraz przypisanej do niego klasy, dzięki czemu zyskuje zdolność klasyfikowania nowych dokumentów wpływających do bazy. Kadhim [39] uzyskał średni stopień zgodności predykcji z wartością rzeczywistą powyżej 80%. W drugim przykładzie uczenie maszynowe jest wykorzystywane do diagnozowanie pacjentów. Cyfrowe bazy danych zawierające ustrukturyzowane informacje o pacjencie, jego objawach oraz zdiagnozowanej chorobie wykorzystywane są do trenowania modeli, które uzyskują zdolności diagnozowania nowych pacjentów [40]. Jan wraz z zespołem [41] uzyskał stopień zgodności predykcji z diagnozą rzeczywistą wynoszący od 92% do 98%. Posiadanie odpowiednich baz danych skraca proces badawczy oraz znacząco wpływa na jakość otrzymanych wyników [40,42,43]. Zespół badawczy może zwrócić szczególną uwagę na opracowywanie algorytmu oraz optymalizowanie jego predykcji, zamiast poświęcać czas na przygotowanie zbiorów danych.

Modele nadzorowanego uczenia maszynowego przyporządkowuje się do dwóch głównych kategorii: regresji oraz klasyfikacji [44–46]. Modele regresji (liniowej oraz

wielozmiennej) zwracają zmienną zależną, która przyjmuje wartości ze zbioru liczb rzeczywistych. Modele te wykorzystywane są w sytuacjach, w których istotna jest dokładna wartość predykcji [44,45,47,48], np. przewidywanie cen akcji na giełdzie, popytu (wielkości sprzedaży), cen nieruchomości, wartości kursu walutowego. Modele klasyfikacji zwracają zmienną zależną, która przyjmuje wartości dyskretne, czyli przypisuje daną obserwację do klasy lub kategorii. Modele te są stosowane w sytuacjach, w których grupowanie obserwacji jest wystarczające [39,44,49–51], np. ocena czy wiadomość e-mail należy oznaczyć jako spam, ocena zdolności kredytowej wnioskującego, grupowanie konsumentów na bazie ich zachowań, kategoryzacja dokumentów tekstowych.

4.2. Uczenia maszynowego stosowane w inżynierii chemicznej

Uczenie maszynowe może mieć zastosowanie w wielu dziedzinach. Synchronicznie z postępującym rozwojem technologii komputerowej oraz nauk informacyjnych wydzieliła się interdyscyplinarna dziedzina określana jako chemoinformatyka [52], co oznacza, że część badań laboratoryjnych (empirycznych) została zastąpiona symulacjami komputerowymi – badania *in silico* (pol. w krzemie; jako analogia do terminów *in vitro, in vivo*).

4.2.1. Tło historyczne

Pomimo, że uczenie maszynowe jest względnie młodym obszarem nauki, to jego zastosowanie w obszarach inżynierii chemicznej ma długą historię – ponad sześćdziesięcioletnią. Już w 1960 roku na łamach *Science* ukazał się artykuł, referujący wykorzystanie kart perforowanych IBM do odnajdowania i sortowania struktur związków chemicznych [53]. Natomiast w 1969 roku ukazała się praca pokazująca wykorzystanie „maszyn uczących się" oraz modelu uczenia maszynowego w celu określania wzoru cząsteczkowego na podstawie widma ze spektrometru masowego o niskiej rozdzielczości. Propozycja ta stanowiła alternatywę do droższej aparatury wysokorozdzielczej [54]. Dynamiczny rozwój techniki komputerowej oraz jej aplikacyjne możliwości doprowadziły do powstania wielu narzędzi informatycznych skierowanych do rozwiązywania problemów inżynierii chemicznej, np. opracowywanie ścieżek syntezy związków organicznych [55]:

- SYNCHEM (rozwinięty w latach 70 XX wieku, Uniwersytet w Nowym Jorku);
- IGOR (rozwinięty w latach 80 XX wieku, Uniwersytet Techniczny w Monachium);
- CHIRON (rozwinięty w latach 80 XX wieku, Uniwersytet Montrealski).

Na przełomie lat osiemdziesiątych i dziewięćdziesiątych ubiegłego wieku były już dostępne prace przeglądowe omawiające dostępne modele, oprogramowanie i narzędzia, które mogą być wykorzystane w pracach optymalizacyjnych i badawczych [56–60].

Szerokie zastosowanie informatyki (nauk informacyjnych) w inżynierii chemicznej sprawiło, że nowy interdyscyplinarny obszar nauki został zdefiniowany – chemoinformatyka. Termin ten pojawia się w literaturze po raz pierwszy w 1998 roku w pracy Browna, który opisuje wpływ nauk informacyjnych w badaniach nad odkrywaniem nowych leków [52]. Od tamtego czasu zakres chemoinformatyki znacząco się zwiększył i nie obejmuje już jedynie badań nad lekami, ale właściwie wszystkie dyscypliny chemiczne od badań podstawowych aż po przemysłowe projekty aplikacyjne [61–65]. Powszechnie stosowana definicja tej dziedziny opracowana została przez Johanna Gasteigera – chemoinformatyka to ogólne użycie informatyki do rozwiązywania problemów chemicznych [65]. Dziedzina ta posiada często cytowane recenzowane czasopisma, takie jak: *Journal of Chemical Information and Modeling* (IF_{2021} = 6,162 [66]), Journal of Cheminformatics (IF_{2020} = 5,514 [67]), *Molecular Informatics* (IF_{2021} = 4,050 [68]). Najnowsza literatura służy badaczom już nie tylko przykładami zastosowania uczenia maszynowego w chemii, czy pracami przeglądowymi, ale również artykułami, które omawiają wytyczne, współcześnie najlepsze praktyki wykorzystywane w tego rodzaju badaniach oraz przykłady projektów realizowanych w przemyśle [69–73].

4.2.2. Przykłady zastosowania uczenia maszynowego

Zastosowanie uczenia maszynowego jest obecnie przedmiotem intensywnych badań, liczba dostępnych publikacji (z dnia 16 kwietnia 2023 roku wyszukiwarka Google Scholar dla frazy „machine+learning+chemistry” zwraca ponad dwa i pół miliona pozycji ogółem, z czego niespełna dwadzieścia tysięcy jest datowanych na ostatnie pięć lat). Sprawia to, że w celach przeglądowych, możliwe jest jedynie wyłonienie kilku głównych obszarów badawczych stosowanych w technologii i inżynierii chemicznej [71]:

a) Opracowywanie nowych związków, materiałów oraz leków.

Projektowanie leków oraz materiałów charakteryzuje się bardzo podobnym profilem zagadnień, które muszą być uwzględnione podczas opracowywania danego produktu. Obie dziedziny uwzględniają: właściwości fizykochemiczne związków, przemysłowy proces produkcji oraz warunki użytkowania produktu. Wszystkie te aspekty są wspierane przez nowoczesne narzędzia modelowania komputerowego [74,75].

Uczenie maszynowe jest stosowane na etapie projektowaniu związków chemicznych w celu przyspieszenia wyboru odpowiedniego indywiduum chemicznego, optymalizacji wyników oraz projektowania technologii wytwarzania. Modelowanie komputerowe wspiera identyfikację związków chemicznych o określonych właściwościach, takich jak aktywność chemiczna i biologiczna, stabilność, rozpuszczalność, toksyczność, struktura przestrzenna [76]. Uczenie maszynowe znajduje zastosowanie w określaniu właściwości związków chemicznych na podstawie ich struktury chemicznej, oddziaływań przestrzennych pomiędzy atomami, czy obecności grup funkcyjnych itp. [77–79] W tym celu modele uczenia maszynowego są trenowane na zbiorach danych zawierających informacje o strukturze chemicznej i właściwościach związków chemicznych [80]. Narzędzia informatyczne zyskują uznanie naukowców, ponieważ pozwalają na redukcję długości trwania projektów badawczych oraz ich kosztochłonności, poprzez: zawężanie liczby związków, które należy poddać laboratoryjnym protokołom badawczym, predykcję właściwości fizykochemicznych rozpatrywanych związków chemicznych, symulację wyników eksperymentów i procedur laboratoryjnych [78,81–83], a co zaś szczególnie tyczy się leków – redukcję liczby badań z wykorzystaniem organizmów żywych [84–89]. Działania te skracają proces opracowania technologii produkcji.

Kolejnym zagadnieniem badawczym, w którym wykorzystywane są wspomniane rozwiązania uczenia maszynowego jest konwersja i przechowywanie energii (opracowywanie baterii). Bateria jest złożonym połączeniem materiałów, a jej właściwości, w tym skuteczność, zależą m.in. od takich procesów jak odwracalne reakcje chemiczne i transport ładunku pomiędzy fazami [90]. Dynamicznie rozwijający się rynek samochodów elektrycznych oraz elektroniki konsumenckiej kreuje zapotrzebowanie na efektywne przechowywanie energii elektrycznej, w związku z czym zespoły badawcze na całym świecie pracują nad usprawnieniem ogniw. Ograniczenie kosztów oraz długości badań laboratoryjnych znacznie zwiększa prawdopodobieństwo komercjalizacji wyników [82,91,92], a umożliwia to szybsze opracowanie produktów dla konsumentów oraz zwiększa konkurencyjność cenową.

Inżynieria chemiczna, a w szczególności tworzenie procesów produkcyjnych opiera się na modelowaniu oraz testowaniu rozwiązań, które służą zaprojektowaniu instalacji produkcyjnej, zapewnieniu jakości produktu, zapewnieniu bezpieczeństwa i higieny pracy oraz rentowności produkcji. Dostępność baz danych oraz narzędzi przewidzianych procesom produkcyjnym powoduje, że modele uczenia maszynowego zyskują przewagę (w kwestiach elastyczności, dokładności, szybkości) nad tradycyjnymi metodami modelowania matematycznego i eksperymentowania [60,93,94]. A w uruchomionych fabrykach algorytmy sztucznej inteligencji wspierają optymalizację procesów m.in. przez sugerowanie optymalnych

nastaw aparatury produkcyjnej (temperatury, ciśnienia, składu surowcowego, wartości przepływów itp.) tak, aby uzyskać maksymalizację takich wskaźników jak wydajność, jakość oraz rentowność [95–98].

b) Planowanie syntez związków oraz optymalizacja reakcji.

Narzędzia oraz techniki uczenia maszynowego wspomagają inżynierów chemików w planowaniu ścieżek reakcyjnych. Ideą tego typu rozwiązań jest to, że program na podstawie wprowadzonej informacji o cząsteczce (np. wzór strukturalny lub notacja SMILES) przedstawia proponowany schemat syntezy wraz ze szczegółami poszczególnych etapów (m.in. potrzebne reagenty, warunki reakcji, procedury). Bazy danych takie jak *SciFinder* oraz *Reaxys* zawierające informacje o milionach reakcji chemicznych, stanowią źródło danych do opracowywania algorytmów uczenia maszynowego tworzących schematy syntez [99,100]. Uproszczone metody zapisywania przebiegu reakcji chemicznej, takie jak notacja Lewisa, są wygodne do stosowania przez człowieka, jednak nie przedstawiają one pełnego obrazu rzeczywistości. Ponadto rosnąca ilość informacji na temat czynników jakie mają wpływ na przebieg reakcji chemicznych sprawia, że pełna analiza i uwzględnienie wszystkich parametrów jest nie do osiągnięcia przez człowieka. Problem ten jest niwelowany przez zastosowanie narzędzi uczenia maszynowego, które są w stanie uwzględnić wszystkie dane dostępne w naukowych bazach danych oraz przeanalizować je w krótkim czasie [101–104]. Analiza danych jest również wykorzystywana do optymalizacji badanej reakcji przez odpowiedni dobór jej parametrów [105,106]. Klasyczną metodą doboru optymalnych warunków reakcji jest iteracyjna zmiana jednego parametru oraz analiza otrzymanych wyników, następnym krokiem jest analiza kolejnego parametru reakcji – metoda ta jest pracochłonna oraz kosztochłonna. Obecnie istnieją algorytmy uczenia maszynowego, które wspierają inżynierów chemików w optymalizowaniu reakcji poprzez predykcje potencjalnie najlepszych warunków [106–108].

c) Interpretacja danych pomiarowych.

W celu zrozumienia oraz odkrywania zjawisk fizykochemicznych, naukowcy wykorzystują wiele różnego rodzaju aparatów badawczych. Nowoczesne techniki takie jak: mikroskopia elektronowa (EM), mikroskopia sił elektronowych (AFM), spektroskopia mas (MS), spektroskopia UV-VIS, magnetyczny rezonans jądrowy (NMR), techniki wykorzystujące promieniowanie rentgenowskie, chromatografia, dostarczają ogromnej ilości informacji. Duże wielowymiarowe zbiory danych będące wynikiem takich badań muszą zostać

przetworzone i przeanalizowane, aby nadać tym danym znaczenie, są to m.in. wyniki odpowiadają danym właściwościom fizykochemicznym (np. pochłanianie promieniowania elektromagnetycznego w spektroskopii) lub parametry identyfikacyjne związków chemicznych (np. czas retencji w chromatografii). Wytrenowane algorytmy uczenia maszynowego wspomagają naukowców w identyfikacji obserwowanych zjawisk i znacznie przyśpieszają proces analizy [109]. Narzędzia uczenia maszynowego, mogą prowadzić do pełniejszego zrozumienia otrzymywanych danych, a więc do maksymalizacji wiedzy o procesach, interakcjach i cechach próbki [110].

Puthongkham wraz z zespołem [111] przedstawia zastosowanie uczenia maszynowego (klasyfikacji oraz regresji) w analizach elektrochemicznych, w celu powiązania wyników pomiarowych (właściwości elektrochemicznych) ze strukturą chemiczną badanego związku. Taka metoda jest stosowana np. w ramach badań nad „elektronicznym językiem", który mógłby zastąpić ocenę sensoryczną wykonywaną przez wykfalifikowanych specjalistów w zakładach przemysłu spożywczego [112–114]. Przykład ten obrazuje również zapotrzebowanie przemysłu w rozwiązania interdyscyplinarne.

Uczenie maszynowe, które wspomaga analizę danych otrzymywanych w technice magnetycznego rezonansu jądrowego (NMR), ma na celu zaproponowanie struktury badanego związku [115,116]. Raljević wraz z zespołem [117] opracował narzędzia, które na bazie danych pochodzących z analiz NMR przewiduje stabilność ropy naftowej. Wydobywane paliwo jest roztworem koloidalnym, który zawiera wiele frakcji. Podczas przetwarzania ropy naftowej mogą tworzyć się agregaty oraz osady, które powodują poważne problemy w procesach przetwórczych, transporcie i magazynowaniu. A zatem informacja o stabilności próbki jest kluczowa dla zakładów produkcyjnych. Do tego momentu nie istniała jedna metoda, która oceniałaby stabilność wszystkich frakcji ropy naftowej.

Również w technikach mikroskopii elektronowej oraz spektroskopii z wykorzystaniem promieniowania rentgenowskiego istnieje wiele zastosowań algorytmów uczenia maszynowego. Obrazy mikroskopowe oraz obrazy natężeń widmowych pochodzących z wielu punktów próbkowania mogą przedstawiać różne składniki chemiczne i ich zależności fizyczne (np. przestrzenne) – jest to bogate źródło informacji o badanym materiale. Interpretacja takich danych wymaga zaawansowanej wiedzy oraz zastosowania przeliczeń statystycznych. Uczenie maszynowe jest zdolne do klasyfikacji danego materiału oraz dostarczyć informację o morfologii danej próbki (np. kształt, struktura, rozmiar) [117–127].

Techniki chromatograficzne (szeroko stosowane w przemyśle chemicznym) w połączeniu z odpowiednim detektorem są w stanie dostarczyć informację o składzie badanej

próbki [128]. Uczenie maszynowe w tej technice ma dwa zastosowania do: predykcji czasu retencji analitu oraz oznaczania molekuł na bazie chromatogramu. Prognozowanie czasu retencji ma na celu dobranie odpowiedniej techniki oraz jej parametrów (m.in. typ fazy stacjonarnej, skład fazy mobilnej) tak by uzyskać pożądany stopień rozkładu analitów. Im bardzie jest złożony skład analizowanej próbki, tym trudniejsze jest to zadanie [129]. Algorytmy uczenia maszynowego z sukcesem są w stanie analizować zbiory danych, dostarczając tym samym informacji o prognozowanym czasie retencji, tak więc iteracyjny eksperymentalny dobór parametrów metody jest znacznie zawężony [130–133]. Problem z oznaczeniem związków na bazie danych pochodzących z chromatografu co do idei jest analogiczny to tego napotykanego w innych technikach analitycznych – duża ilość danych oraz nakładanie się sygnałów pochodzących od analitu. Uczenie maszynowe ułatwia identyfikację związków na bazie chromatogramu (analiza jakościowa), ale także wspomaga oznaczenia ilościowe danych związków (np. usprawnianie integracji pików) [134–137].

4.3. Metody klasyfikacji binarnej

Obecnie jest dostępny szereg narzędzi, które pozwalają wykorzystywać uczenie maszynowe. Model może być budowany od podstaw przy użyciu powszechnych języków programowania, takich jak *Python, C++, Java* [138]. Do dyspozycji są również aplikacje (programy) z wbudowanymi podstawowymi modelami, które mogą być parametryzowane na potrzeby danego projektu, np. *R-Project* [139,140], *MatLab* [141], *TensorFlow* [142]. Istotny jest jednak fakt, że uczenie maszynowe to przeliczenia matematyczne, które są wykonywane w zdefiniowany sposób (algorytmicznie), a więc każde narzędzie zdolne do wykonania potrzebnych operacji może służyć do implementacji takich algorytmów [143,144].

Modele klasyfikacyjne odnoszą się do problemów, w których daną obserwację przypisuje się do odpowiedniej klasy lub kategorii. Szczególnym typem jest klasyfikacja binarna, która przyporządkowuje obserwację do jednej z dwóch klas (np. prawda lub fałsz) [145].

4.3.1. Przykłady algorytmów klasyfikacji binarnej

W literaturze przedmiotu najczęściej występują pięć algorytmów wykorzystywanych do implementacji klasyfikacji binarnej [146]:

- regresja logistyczna (ang. *logistic regression classifier*);
- k-najbliższych sąsiadów (ang. *k-nearest neighbors classifier*);

– drzewa klasyfikacyjne (ang. *decision tree classifier*);
– losowy las decyzyjny (ang. *random forest classifier*);
– naiwny klasyfikator Bayesa (ang. *naïve Bayes classifier*);

4.3.2. Algorytm typu regresji logistycznej

Algorytm regresji logistycznej jest modelem wykorzystywanym do klasyfikacji binarnej. Jest to model liniowy nadzorowanego uczenia maszynowego, którego podstawowym założeniem jest to, że obserwacje (dane) mogą być rozdzielone poprzez linię lub płaszczyznę *n*-wymiarową. Predykcja klasy następuje poprzez wyliczenie wartości wielomianu pierwszego stopnia zgodnie ze wzorem (1).

$$y = \omega_1 x_1 + \omega_2 x_2 + \cdots \omega_n x_n \tag{1}$$

Zmienna *x* jest zmienną niezależną i oznacza parametry jakie są uwzględnianie w celu wyliczania predykcji, *ω* oznacza wagę przypisaną towarzyszącemu parametrowi, *n* zaś stanowi liczbę parametrów. Ponieważ jest to szacowanie prawdopodobieństwa, które przybiera wartość z przedziału od 0 do 1, to w kolejnym kroku obliczona wartość *y* jest przeliczana za pomocą funkcji, której wynik zwróci liczbę *y'* z zakresu od 0 do 1 – np. tangens hiperboliczny (wzór (2)) lub funkcja logitowa (wzór (3)).

$$y' = \tanh(y) = \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}} \tag{2}$$

$$y' = \text{logit}(y) = \ln\left(\frac{y}{y-1}\right) \tag{3}$$

Jako że jest to algorytm binarny, to przyjmuje się, że wartości $y' \in (0;\ 0{,}5)$ odpowiadają klasie pierwszej, zaś wartości $y' \in \langle 0{,}5;\ 1)$ odpowiadają klasie drugiej. Może mieć miejsce dostosowywanie punktu podział tego zakresu, aby uzyskać najlepszą odpowiedź [147,148].

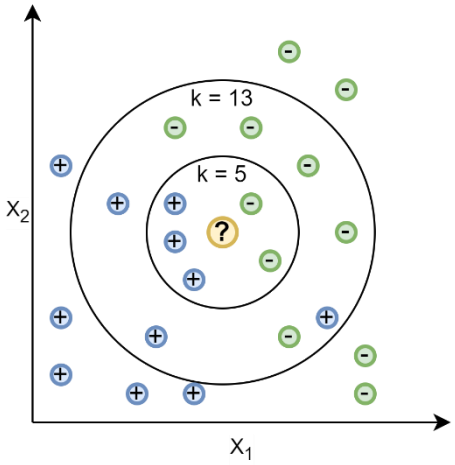
Proces uczenia algorytmu polega na analizie zbioru danych, który składa się z par obserwacja-etykieta. Następnie dobierane się wartości wag *ω* tak aby odpowiedzi algorytmu były w jak najwyższym stopniu zgodne z inicjalnie przypisanymi etykietami.

4.3.3. Algorytm typu *k*-najbliższych sąsiadów

Algorytm *k*-najbliższych sąsiadów (ang. *k-nearest neighbors*, kNN) polega na klasyfikacji nowej próbki na podstawie jej podobieństwa do obserwacji w zbiorze treningowym – podobieństwo wyrażane jest jako odległość euklidesowa pomiędzy dwoma punktami, obliczana według wzoru (4) [149].

$$\mathrm{d}_z(A, B_z) = \sqrt{\left(x_{1A} - x_{1B_z}\right)^2 + \cdots + \left(x_{nA} - x_{nB_z}\right)^2} \tag{4}$$

Działanie algorytmu składa się z trzech głównych etapów. W pierwszym kroku wyznaczana jest odległość d_z od badanego punktu A do punktu B_z pochodzącego ze zbioru treningowego (wzór (4)). Parametry (zmienne niezależne) oznaczone są jako x, zaś ich liczba jako n. Wartość d_z jest obliczana dla każdej obserwacji B_z, ze zbioru treningowego (z – liczba punktów w bazie danych). Wynikiem tego etapu jest zestaw odległości od $d_1(A, B_1)$ do $d_z(A, B_z)$. W drugim etapie wybierani są najbliżsi sąsiedzi, czyli te punkty, które charakteryzują się najmniejszą wartością odległości. Liczba sąsiadów jest oznaczana współczynnikiem k, który jest ustalany przez użytkownika. Trzecim, ostatnim krokiem jest zaklasyfikowanie obserwacji badanej. Klasa punktu A jest tą, która najczęściej występuje wśród wybranych k-najbliższych sąsiadów. Na Rysunku 4.3 jest przedstawiona graficznie metoda działania algorytmu [149,150].



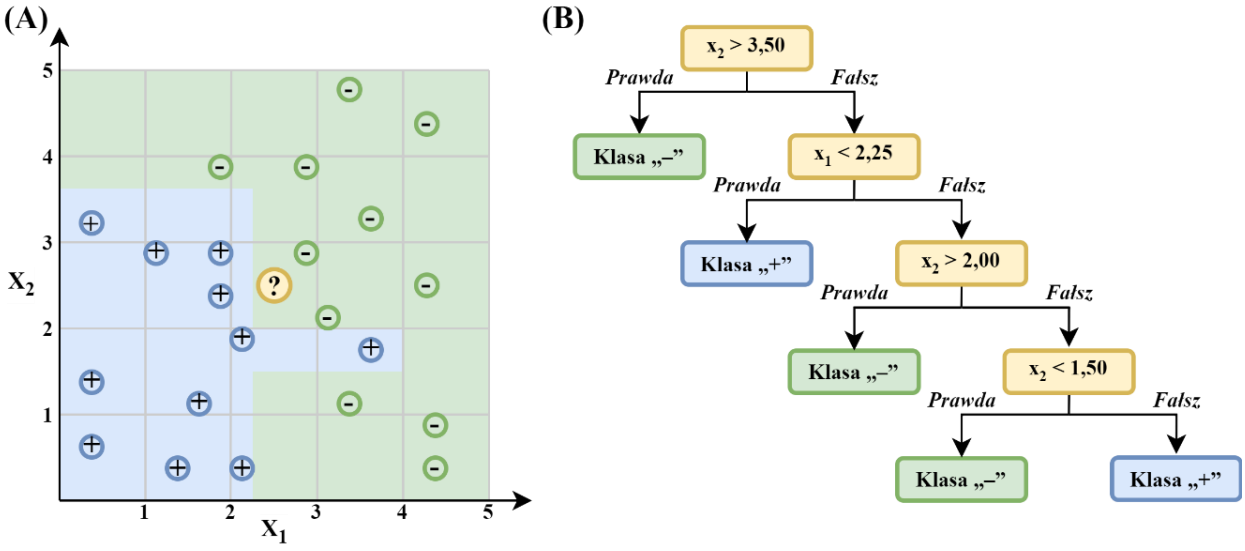
Rysunek 4.3. Graficzna interpretacja działania metody k-najbliższych sąsiadów [149]

Obserwacje przedstawione na Rysunku 4.3. posiadają dwie zmienne niezależne: x_1 oraz x_2. Punkt badany oznaczony jako „?” może zostać przyporządkowana do jednej z dwóch klas: „+” lub „–”. Gdy k = 5 to obserwacji badanej zostanie przypisana klasa „+”, ponieważ największa liczba sąsiadów posiada tą klasę (3 sąsiadów z 5). Jeżeli jednak k = 13, to na postawie tego samego zbioru punktów, przypisaną klasą będzie „–” (7 sąsiadów z 13).

4.3.4. Algorytm typu drzewa klasyfikacyjne

Rodzina metod analizy danych, które wizualizuje się za pomocą diagramów decyzyjnych określa się mianem drzew decyzyjnych. Szczególnym przypadkiem są drzewa klasyfikacyjne, których celem jest przyporządkowanie obserwacji do odpowiedniej grupy. Według powszechnej nomenklatury drzewo składa się z węzłów, gałęzi oraz liści. Warunki (węzły) są tworzone w taki sposób, aby odpowiedź na nie była prawdą bądź fałszem. W przypadku małych zbiorów danych struktura drzewa (łącznie z warunkami) może być tworzona przez człowieka. Modele uczenia maszynowego analizujące duże bazy danych wykorzystują w tym celu wiele teorii statystycznych np. mierzenie losowość i zanieczyszczenia zbioru danych za pomocą entropii. Pomimo intuicyjnie logicznej struktury to aparat matematyczny nie jest jednolity (wykorzystuje wiele teorii). Istnieją wyspecjalizowane narzędzia informatyczne, które służą do modelowanie uczenia maszynowego z wykorzystaniem drzew klasyfikacyjnych [50,151].

Na Rysunku 4.4 przedstawiono wizualizację tworzenia oraz działania algorytmu drzewa klasyfikacyjnego.

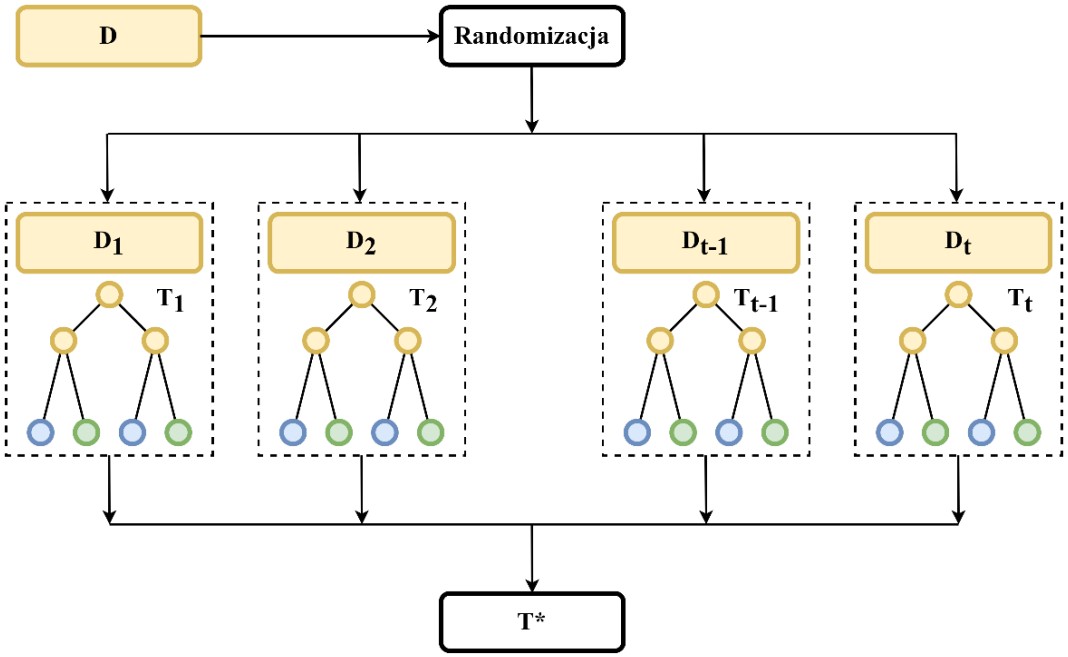


Rysunek 4.4. Wizualizacja algorytmu drzewa klasyfikacyjnego; (A) zbiór danych treningowych; (B) przykładowe drzewo klasyfikacyjne.

Na Rysunku 4.4A przedstawione są punkty treningowe z przypisanymi im klasami „+” lub „–”. Każdy punkt posiada dwie zmienne niezależne x_1 oraz x_2. Drzewo klasyfikacyjne (Rysunek 4.4B) jest tworzone w taki sposób, by wydzielić obszary odpowiadające występowaniu każdej z klas – zawiera ono 4 węzły. Istnieje niepoliczalna liczba sposobów, według których możliwe jest wyodrębnienie tych obszarów. Obserwacji badanej oznaczonej jako „?” na Rysunku 4.4A przypisuje się klasę „–” (należy odpowiedzieć na 3 warunki).

4.3.5. Algorytm typu losowy las decyzyjny

Losowe lasy decyzyjne należą do grupy technik, które wykorzystują zespoły klasyfikatorów. Takie rozwiązanie znajduje zastosowanie w analizie obszernych zbiorów danych posiadających wiele wymiarów (paramentów). W przypadku tego typu algorytmu wielokrotnie używana jest teoria drzewa klasyfikacyjnego (opisana na stronie 32). Oznacza to, że w trakcie tworzenia oraz trenowania modelu uczenia maszynowego, diagram decyzyjny jest tworzony wielokrotnie. To rozwiązanie cechuje się wysoką precyzją i dokładnością. Znajduje ono zastosowanie w medycynie jak i w bankowości. Rekomendowane jest do zbiorów danych, które intensywnie się rozszerzają i/lub mogą zawierać braki w informacjach. Analogicznie do pojedynczego drzewa decyzyjnego, aparat matematyczny jest skomplikowany, a zastosowanie bez odpowiedniego oprogramowania mozolne. Na Rysunku 4.5 przedstawiona została graficzna reprezentacja algorytmu [152,153].



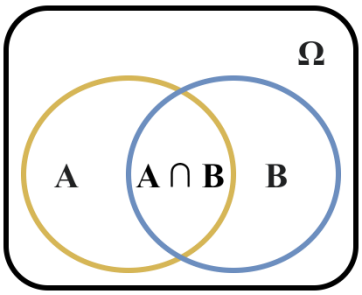
Rysunek 4.5. Reprezentacja algorytmu typu losowy las decyzyjny [153].

Zgodnie z przedstawionym schematem (Rysunek 4.5) tworzenie modelu uczenia maszynowego rozpoczyna się od losowego podziału danych treningowych *D* – proces ten określany jest randomizacją. Wynikiem tego etapu są wyodrębnione podzbiory od D_1 do D_t. W kolejnym kroku do każdego podzbioru tworzone jest drzewo decyzyjne od T_1 do T_t. Klasyfikacja nowej obserwacji następuje „przez głosowanie”, oznacza to, że zostaje przypisana ta klasa (*T**), do której prowadziła największa liczba drzew.

4.3.6. Naiwny klasyfikator Bayesa

Prawdopodobieństwo warunkowe (twierdzenia Bayesa)

Naiwne klasyfikatory Bayesa opierają się na teorii prawdopodobieństwa warunkowego – twierdzenie Bayesa. Na Rysunku 4.6 przedstawiono ideę prawdopodobieństwa całkowitego, w którym wydzielone są dwa podzbiory zdarzeń posiadające część wspólną [154].



Rysunek 4.6. Diagram przedstawiający teorię całkowitego prawdopodobieństwa z wydzielonymi podzbiorami A oraz B [154].

Na Rysunku 4.6 wydzielony obszar Ω zawiera wszystkie możliwe zdarzania. Wydzielone zostały również zdarzenia *A* oraz *B*. Obszar, w którym występuje zarówno zdarzenie *A* oraz *B* jest iloczynem tych dwóch podzbiorów ($A \cap B$). Prawdopodobieństwo warunkowe *P(A|B)* to prawdopodobieństwo wystąpienia zdarzenia *A* pod warunkiem zaistnienia zdarzenia *B*, wyznaczane jest na podstawie twierdzenia Bayesa – wzór (5) [154,155]:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)} \tag{5}$$

gdzie:

P(A|B) – prawdopodobieństwo wystąpienia *A* pod warunkiem, że zaistniało *B*,

P(B|A) – prawdopodobieństwo wystąpienia *B* pod warunkiem, że zaistniało *A*,

P(A) – prawdopodobieństwo wystąpienia zdarzenia *A*,

P(B) – prawdopodobieństwo wystąpienia zdarzenia *B*.

Zastosowanie twierdzenia Bayesa w algorytmie klasyfikującym

Zadaniem klasyfikatorów jest przyporządkowanie badanej obserwacji do jednej ze zdefiniowanych kategorii. Innymi słowy jest to prawdopodobieństwo warunkowe zaistnienia danej klasy, gdzie warunkiem jest wystąpienie obserwacji badanej. Algorytm uczenia maszynowego realizowany jest poprzez sprawdzanie dla której z kategorii wyznaczona została największa wartość na podstawie twierdzenia Bayesa – wzór (6). W klasyfikacji binarnej spomiędzy dwóch klas, do badanej obserwacji zostaje przypisana ta, dla której zostało wyznaczone większe prawdopodobieństwo warunkowe $P(C_n|X)$. Dla każdej z klas prawdopodobieństwo wystąpienia obserwacji $P(X)$ będzie stałe, tak więc dla algorytmu klasyfikacji istotny jest jedynie licznik (wzór (6)) [154,155].

$$P(C_n|X) = \frac{P(X|C_n)\,P(C_n)}{P(X)} \tag{6}$$

gdzie:

$P(C_n|X)$ – prawdopodobieństwo wystąpienia C_n pod warunkiem zaistnienia obserwacji X,

C_n – klasa o indeksie n,

n – liczba klas,

$P(X|C_n)$ – prawdopodobieństwo wystąpienia X pod warunkiem zaistnienia klasy C_n,

$P(C_n)$ – prawdopodobieństwo wystąpienia klasy C_n,

$P(X)$ – prawdopodobieństwo wystąpienia obserwacji X.

Badana obserwacja X może posiadać wiele zmiennych niezależnych od x_1 do x_i (i oznacza liczbę argumentów). Klasyfikator nazywany jest „naiwnym" ponieważ zakłada, że argumenty składowe (x_i) są niezależne od siebie. Dlatego możliwe jest przyjęcie, że prawdopodobieństwo całego wektora X jest iloczynem prawdopodobieństw swoich składowych. A więc na potrzeby algorytmu klasyfikującego stosowany jest wzór (7):

$$P(X|C_n) = P(x_1|C_n) \cdot \ldots \cdot P(x_i|C_n) = \prod_1^i P(x_i|C_n) \tag{7}$$

Z wykorzystanie danych historycznych, w których każdy punkt posiada przypisaną klasę (etykietę) możliwe jest obliczenie poszczególnych prawdopodobieństw używanych we wzorach (6) i (7) [154–156].

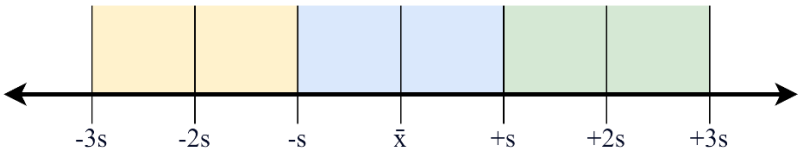
Dyskretyzacja zmiennych

Naiwny klasyfikator Bayesa może być wykorzystywany zarówno dla zmiennych ciągłych (wartości rzeczywiste, np. pH od 1 do 14) jak i dyskretnych (liczba wartości jakie przyjmuje zmienna są skończone lub policzalne, np. pH zasadowe, obojętne, kwasowe). Modele uczenia maszynowego oparte na zmiennych dyskretnych charakteryzują się większą dokładnością i odpornością na błędy. Bardzo często w świecie naturalnym zmienne przybierają wartości liczb rzeczywistych. Jeżeli takie parametry mają być wykorzystywane w modelu uczenia maszynowego, aby zwiększyć jego efektywność, należy zastosować proces dyskretyzacji zmiennych. Polega on na transformacji wartości ze zbioru liczb rzeczywistych na jej dyskretny odpowiednik – przedział. Istnieje szereg technik, które mogą służyć temu celowi. Najprostszą z nich jest podział zakresu wartości zmiennej na skończoną liczbę równych przedziałów (ang. *Equal Width Discretization*, EWD) [156–159], zgodnie ze wzorem (8):

$$EWD = \frac{x_{max} - x_{min}}{k} \tag{8}$$

Szerokości przedziału EWD rozpatrywanej zmiennej x jest określana na podstawie jej wartości maksymalnej x_{max} oraz wartości minimalnej x_{min}. Parametr k stanowi liczbę przedziałów na jaką dany zakres ma być podzielony.

Inną metodą, która może być wykorzystana do wydzielenia odpowiednich przedziałów (przy założeniu, że zmienna ma rozkład normalny), jest dyskretyzacja oparta o odchylenie standardowe [159–161]. W tym przypadku, granice przedziałów są uzależnione od wartości odchylenia standardowego (Rysunek 4.7).



Rysunek 4.7. Graficzna interpretacja dyskretyzacji opartej o odchylenie standardowe [160].

W przykładzie przedstawionym na Rysunku 4.7 zmienna rzeczywista x została poddana dyskretyzacji na trzy przedziały. Ich szerokość została ustalona na wartość równą dwóm odchyleniom standardowy ($2s$). Tak więc przedział wyśrodkowany względem wartości średniej $\bar{x}$ zawiera się w zbiorze liczb $\langle -s;\ +s)$.

Przykład zastosowania naiwnej klasyfikacji Bayesa

Zastosowanie naiwnego klasyfikatora Bayesa można zademonstrować na wielu przykładach pochodzących z życia codziennego np. ocena zdolności kredytowej, prognozowanie pogody, oznaczanie e-maili jako niechciane.

Poniżej zaprezentowano zastosowanie klasyfikacji, która ma na celu odpowiedzieć na pytanie czy odczynnik powinien być oczyszczony przed użyciem. Laboratorium posiada informację o historycznych partiach substratu, o cenie oraz czy wymagane było wcześniejsze oczyszczanie substancji (Tabela 4.1).

Tabela 4.1. Dane ilustracyjne do przykładu zastosowania naiwnej klasyfikacji Bayesa

Liczba dni od otwarcia	Procent zużycia	Użyty przez studenta	Koszt	Wymagane oczyszczanie
50	20	Tak	Wysoki	Tak
40	70	Tak	Wysoki	Tak
18	75	Tak	Wysoki	Nie
11	18	Tak	Niski	Nie
53	15	Nie	Niski	Nie
35	70	Nie	Niski	Tak
45	85	Tak	Wysoki	Nie
65	30	Nie	Niski	Nie

W Tabeli 4.1 zestawione są zmienne o wartościach rzeczywistych (liczba dni od otwarcia, procent zużycia), parametry dyskretne (użyty przez student, koszt) oraz etykieta (wymagane oczyszczanie). Są to dane treningowe. Tak więc pierwszym etapem tworzenia klasyfikacji jest dyskretyzacja zmiennych ciągłych. Wiek odczynnika został podzielony na trzy grupy: grupa 1 (mniejszy równy 30 dni), grupa 2 (od 31 do 50 dni), grupa 3 (powyżej 50 dni). Zaś procent zużycia został zdefiniowany jako niski (poniżej 40%), średni (od 40% do 70%) oraz wysoki (powyżej 70%). Dane po transformacji przedstawione są w Tabeli 4.2.

Tabela 4.2. Dane ilustracyjne po dyskretyzacji do przykładu naiwnej klasyfikacji Bayesa.

Liczba dni od otwarcia	Procent zużycia	Użyty przez studenta	Koszt	Wymagane oczyszczanie
Grupa 2	Niski	Tak	Wysoki	Tak
Grupa 2	Średni	Tak	Wysoki	Tak
Grupa 1	Wysoki	Tak	Wysoki	Nie
Grupa 1	Niski	Tak	Niski	Nie
Grupa 3	Niski	Nie	Niski	Nie
Grupa 2	Średni	Nie	Niski	Tak
Grupa 2	Wysoki	Tak	Wysoki	Nie
Grupa 3	Niski	Nie	Niski	Nie

Odczynnik (obserwacja badana), który jest poddawany klasyfikacji charakteryzuje się danymi zestawionymi w Tabeli 4.3.

Tabela 4.3. Dane ilustracyjne obserwacji badanej (odczynnika) poddawanej klasyfikacji.

Liczba dni od otwarcia	Procent zużycia	Użyty przez studenta	Koszt
45	16	Tak	Wysoki

W celu dokonania klasyfikacji informacje o obserwacji badanej (Tabela 4.3) należy poddać takiej samej dyskretyzacji jak wartości historyczne z Tabela 4.1 (zbiór treningowy). Parametry substratu po transformacji zestawione są w Tabeli 4.4.

Tabela 4.4. Dane ilustracyjne po dyskretyzacji obserwacji badanej poddawanej klasyfikacji.

Liczba dni od otwarcia	Procent zużycia	Użyty przez studenta	Koszt
Grupa 2	Niski	Tak	Wysoki

Odczynnik może zostać przypisany do jednej z dwóch klas C_{WO} (wymaga oczyszczenia) oraz C_{NWO} (nie wymaga oczyszczenia). Oznacza się prawdopodobieństwo wystąpienia każdej z klas w zbiorze danych treningowych (Tabela 4.2) – równania (A0) oraz (B0).

$$P(C_{WO}) = \frac{3}{8} = 0{,}375 \quad \text{(A0)} \qquad P(C_{NWO}) = \frac{5}{8} = 0{,}625 \quad \text{(B0)}$$

Następnie dla każdej wartości argumentu obserwacji badanej (Tabela 4.4) oznacza się prawdopodobieństwa warunkowe jej wystąpienia w zbiorze treningowym (Tabela 4.2), gdzie warunek stanowi każda z klas – równania od (A1) do (A4) oraz od (B1) do (B4).

$$P(Wiek = \ grupa\ 2|C_{WO}) = \frac{3}{3} = 1 \quad \text{(A1)} \qquad P(Wiek = \ grupa\ 2|C_{NWO}) = \tfrac{1}{5} = 0{,}2 \quad \text{(B1)}$$

$$P(Zużycie = niskie|C_{WO}) = \frac{1}{3} \approx 0{,}333 \quad \text{(A2)} \qquad P(Zużycie = niskie|C_{NWO}) = \frac{3}{5} = 0{,}6 \quad \text{(B2)}$$

$$P(Student = tak|C_{WO}) = \frac{2}{3} \approx 0{,}667 \quad \text{(A3)} \qquad P(Student = tak|C_{NWO}) = \frac{3}{5} = 0{,}6 \quad \text{(B3)}$$

$$P(Koszt = \ wysoki|C_{WO}) = \frac{2}{3} \approx 0{,}667 \quad \text{(A4)} \qquad P(Koszt = \ wysoki|C_{NWO}) = \frac{2}{5} = 0{,}4 \quad \text{(B4)}$$

W kolejnym kroku wyznacza się prawdopodobieństwa warunkowego całego wektora. Zgodnie ze wzorem (7) jest to iloczyn wartości prawdopodobieństw składowych. W omawianym przykładzie wektor *X* składa się z czterech argumentów: wieku odczynnika, procentu zużycia, użycia przez studenta oraz kosztu. W równaniu (A5) obliczana jest wartość prawdopodobieństwa wystąpienia obserwacji pod warunkiem klasy pierwszej – $P(X|C_{WO})$. Jest ono równe iloczynowi wartości otrzymanych w równaniach od (A1) do (A4).

$$P(X|C_{WO}) = 1 \cdot 0{,}333 \cdot 0667 \cdot 0{,}667 \approx 0{,}148 \tag{A5}$$

Odpowiednio dla klasy drugiej (równanie (B5)) wartość prawdopodobieństwa $P(X|C_{NWO})$ stanowić będzie iloczyn wartości z równań od (B1) do (B4).

$$P(X|C_{NWO}) = 0{,}2 \cdot 0{,}6 \cdot 0{,}6 \cdot 0{,}4 \approx 0{,}029 \tag{B5}$$

Ostatnim krokiem jest sprawdzenie, która z klas posiada większe prawdopodobieństwo zgodnie ze wzorem (6). W przypadku klasyfikacji binarnej jest to sprawdzenie, która z dwóch wartości jest większa. Dzięki temu, że ważna jest jedynie relacja pomiędzy wartościami, możliwe jest pominięcie mianownika ze wzoru (6). Tym samym otrzymuje się równanie (C0):

$$\begin{aligned} max\big(P(C_n|X)\big) &= max\,\{P(C_{WO}|X)\ ;\ P(C_{NWO}|X)\} \\ &= max\left\{\frac{P(X|C_{WO})\,P(C_{WO})}{P(X)}\ ;\ \frac{P(X|C_{NWO})\,P(C_{NWO})}{P(X)}\right\} \\ &= max\,\{P(X|C_{WO})\,P(C_{WO})\ ;\ P(X|C_{NWO})\,P(C_{NWO})\} \end{aligned} \tag{C0}$$

Wykorzystując wartości z równań (A0), (A5), (B0), (B5) i podstawiając je do równania (C0) otrzymuje się końcowy wynik – równanie (C1).

$$\begin{aligned} max\big(P(C_n|X)\big) &= max\{P(X|C_{WO})\,P(C_{WO})\ ;\ P(X|C_{NWO})\,P(C_{NWO})\} \\ &= max\{0{,}375 \cdot 0{,}148\,;0{,}625 \cdot 0{,}029\} \\ &= max\{0{,}056\,;0{,}018\} \\ &= P(C_{WO}|X) \end{aligned} \tag{C1}$$

Wartość prawdopodobieństwa dla klasy C_{WO} (wymaga oczyszczenia) jest większa, niż prawdopodobieństwo dla klasy C_{NWO} (nie wymaga oczyszczenia). Zatem obserwacja badana (odczynnik) należy poddać oczyszczeniu.

4.4. Wybrane metody oceny klasyfikatorów binarnych

Integralną częścią tworzenia narzędzi na bazie uczenia maszynowego jest weryfikacja ich działania. Istnieje wiele wskaźników i technik, które mogą być wykorzystane w tym celu. Odpowiednia ocena dostarcza informacji zwrotnej o efektach trenowania algorytmu oraz sygnalizuje użytkownikowi poziom jakości (skuteczności) prezentowanej klasyfikacji.

Algorytmy klasyfikacji binarnej przypisują obserwację badaną do jednej z dwóch klas. Powszechną notacją jest normalizacja odpowiedzi klasyfikatora do wartości pozytywnej (przypisanie do klasy pierwszej) oraz negatywnej (przypisanie do klasy drugiej). Pozwala to na standaryzację nazewnictwa niezależnie od tego czym są rozpatrywane klasy w rzeczywistości.

4.4.1. Tablica pomyłek

Tablica pomyłek, zwana również macierzą błędów (ang. *confusion matrix*), jest narzędziem stosowanym do oceny jakości klasyfikacji. Pozycje ze zbioru treningowego oznacza się etykietami pozytywną lub negatywną. Punkt otrzymuje nową, predykcyjną etykietę, która jest wynikiem działania klasyfikatora. Wszystkie możliwe sytuacje zgodności predykcji z wartością pierwotną przedstawia tablica pomyłek (Rysunek 4.8) [162–165].

		Wartości rzeczywiste	
		AP Rzeczywiście pozytywne (ang. *actual positive*)	**AN** Rzeczywiście negatywne (ang. *actual negative*)
Wartości prognozowane	**PP** Prognozowane pozytywne (ang. *predicted positive*)	**TP** Prawdziwie pozytywne (ang. *true positive*)	**FP** Fałszywie pozytywne (ang. *false positive*)
	PN Prognozowane negatywne (ang. *predicted negative*)	**FN** Fałszywie negatywne (ang. *false negative*)	**TN** Prawdziwie negatywne (ang. *true negative*)

Rysunek 4.8. Tablica pomyłek klasyfikatora binarnego [163].

Tablica pomyłek przedstawiona na Rysunku 4.8 posiada dwa wiersze i dwie kolumny. Jest ona tworzona poprzez porównanie wartości prognozowanej (wiersze) z wartością rzeczywistą (kolumny). Możliwe jest zaobserwowanej 4 porównań prognozy ze stanem faktycznym (2 dla zgodności i 2 dla niezgodności) [163,165]:

- prawdziwie pozytywne (ang. *true positive,* TP): zgodność predykcji etykiety pozytywnej z rzeczywistą wartością pozytywną w zbiorze treningowym,
- prawdziwie negatywne (ang. *true negative,* TN): zgodność predykcji etykiety negatywnej z rzeczywistą wartością negatywną w zbiorze treningowym,
- fałszywie pozytywne (ang. *false positive,* FP): niezgodność predykcji etykiety pozytywnej z rzeczywistą wartością negatywną w zbiorze treningowym,
- fałszywie negatywne (ang. *false negative*, FN): niezgodność predykcji etykiety negatywnej z rzeczywistą wartością pozytywną w zbiorze treningowym.

Wartości TP, TN, FP, FN bezpośrednio prezentują jakość oceny klasyfikatora. Jednak aby uzyskać pogłębioną analizę na temat skuteczności narzędzia wykorzystuje się wiele innych wskaźników. Podstawę do ich wyliczania stanowią jednak te cztery parametry [162,163].

Na Rysunku 4.8. wskazane są dodatkowo wartości:

- rzeczywiście pozytywne (ang. *actual positive,* AP): liczba próbek ogółem, które w zbiorze treningowym posiadają etykietę pozytywną – wzór (9),
- rzeczywiście negatywne (ang. *actual negative,* AN): liczba próbek ogółem, które w zbiorze treningowym posiadają etykietę negatywną – wzór (10),
- prognozowane pozytywne (ang. *predicted positive*, PP): liczba próbek ogółem, które zostały zaklasyfikowane przez algorytm jako pozytywne – wzór (11),
- prognozowane negatywne (ang. *predicted negative*, PN): liczba próbek ogółem, które zostały zaklasyfikowane przez algorytm jako negatywne – wzór (12).

Wartości te mogą być odczytane bezpośrednio ze zbiorów danych. Możliwe jest wykonie weryfikacji poprzez sprawdzenie czy zachodzą równości zgodnie ze wzorami od (9) do (12).

$$AP = TP + FN \tag{9}$$

$$AN = FP + TN \tag{10}$$

$$PP = TP + FP \tag{11}$$

$$PN = FN + TN \tag{12}$$

4.4.2. Dokładność i poziom błędów

Dokładność (ang. *accuracy*, ACC) jest jedną z najpowszechniej stosowanych miar oceny skuteczności algorytmu. Jest ona definiowana jako stosunek liczby próbek prawidłowo zaklasyfikowanych do liczby próbek ocenionych – obliczana jest zgodnie ze wzorem (13). Przyjmuje wartości od 0 do 1, czyli od najlepszej do najgorszej dokładności [163,165].

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{AP + AN} \tag{13}$$

Miarą komplementarną do dokładności jest poziom błędów (ang. *error rate*, ERR). Jest on definiowany jako stosunek liczby próbek błędnie zaklasyfikowanych do liczby próbek ocenionych – obliczany jest zgodnie ze wzorem (14). Im mniejsza wartość współczynnika tym lepsza odpowiedź algorytmu [163,165].

$$ERR = \frac{FP + FN}{TP + TN + FP + FN} = \frac{FP + FN}{AP + AN} = 1 - ACC \tag{14}$$

W przypadku gdy oceniany algorytm będzie uczony na zbiorze treningowym, w którym klasy nie są równoliczne (zbalansowane) współczynniki te mogę prezentować zbyt ogólną informację. W przypadku wysokiej wartości dokładności w liczniejszej klasie zachodzi równoważenie (maskowanie) niskiej dokładności w klasie mniej licznej – wynik ACC jest prezentowany jako pojedyncza liczba, tak więc przedstawiana wartość może być zafałszowana. Analogiczna sytuacja będzie obserwowana dla poziomu błędów [162,164–166].

4.4.3. Precyzja

Precyzja pozytywna (ang. *positive predicted value*, PPV) jest stosunkiem próbek prawdziwie oznaczonych jako pozytywne do liczby próbek oznaczonych przez algorytm jako pozytywne – wzór (15). Dla obserwacji oznaczonej przez klasyfikator etykietą pozytywną PPV jest prawdopodobieństwem, że jest ona w rzeczywistości pozytywna [163–166].

$$PPV = \frac{TP}{TP + FP} = \frac{TP}{PP} \tag{15}$$

4.4.4. Czułość i specyficzność

Czułość (ang. *true positive rate*, TPR) wyraża stosunek liczby próbek prawidłowo zaklasyfikowanych jako pozytywne do ogólnej liczby próbek pozytywnych w zbiorze treningowym – wzór (16). Wartość współczynnika równa 1 oznacza, że wszystkie próbki oznaczone pierwotnie jako pozytywne zostały poprawnie sklasyfikowane [163,165].

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{AP} \tag{16}$$

Specyficzność (ang. *true negative rate*, TNR) jest komplementarną miarą dla drugiej klasy. Wyraża stosunek liczby próbek prawidłowo zaklasyfikowanych jako negatywne do ogólnej liczby próbek negatywnych w zbiorze treningowym - wzór (17) [163,165].

$$TNR = \frac{TN}{FP + TN} = \frac{TN}{AN} \tag{17}$$

Oba współczynnik są odpowiednikiem dokładności dla każdej z klas z osobna. Zatem TPR oraz TNR są wrażliwe tylko na własną grupę. Używanie tych współczynników zamiast dokładności (ACC) jest szczególnie istotne, kiedy zbiór danych treningowych nie jest zbalansowany – klasy nie są równoliczne [162,163,165,166].

4.4.5. Częstość fałszywych alarmów oraz częstość fałszywych odkryć

Częstość fałszywych alarmów (ang. *false positive rate*, FPR) jest stosunkiem liczby próbek fałszywie oznaczonych jako pozytywne do liczby próbek, które posiadających etykietę negatywną w zbiorze treningowym – wzór (18). Podczas trenowania algorytmu dąży się do minimalizacji tego wskaźnika [163,165].

$$FPR = \frac{FP}{FP + TN} = \frac{FP}{AN} \tag{18}$$

Częstszość fałszywych odkryć (ang. *false negative rate*, FNR) jest analogicznym wskaźnikiem do FPR. Wyznacza on stosunek liczby próbek zaklasyfikowanych jako fałszywie negatywne do liczby próbek pierwotnie oznaczonych jako pozytywne – wzór (19) [163,165].

$$FNR = \frac{FN}{FN + TP} = \frac{FP}{AP} \tag{19}$$

Te wskaźniki, tak samo jak TPR oraz TNR, są wrażliwe tylko na odpowiadającą im klasę, zatem są stosowane w przypadku niezbalansowanej bazy danych treningowych [162,163,165].

4.4.6. Współczynnik korelacji Matthewsa

Współczynnik korelacji Matthewsa (ang. *Matthews correlation coefficient*, MCC) wprowadzony został przez biochemika Briana W. Matthewsa w 1975 roku. Jest miarą jakości klasyfikacji binarnych (dwuklasowych) w uczeniu maszynowym. MCC bierze pod uwagę wszystkie cztery przypadki zgodności pierwotnej etykiety z predykcją klasyfikatora – wzór (20). Od dokładności (ACC) odróżnia go szczególnie to, że uważany jest za zrównoważoną miarę, która może być używana nawet jeśli klasy są bardzo różnej liczności. Kolejną istotną kwestią jest to, że miara ta jest niezależna od tego, która klasa zostanie oznaczona jako pozytywna. MCC jest w istocie współczynnikiem korelacji między etykietami rzeczywistymi (ze zbioru treningowego) i przewidywanymi przez klasyfikator. Przybiera wartości od +1 do –1, gdzie: +1 oznacza idealne przewidywanie, 0 reprezentuje predykcję nie lepszą niż losową, zaś –1 opisuje całkowitą niezgodność pomiędzy wynikiem klasyfikacji, a pierwotnymi etykietami [163,164,166].

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{20}$$

4.4.7. Współczynnik F_1-score

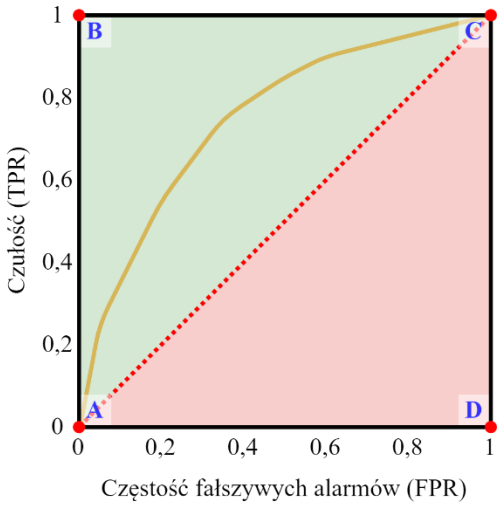
Współczynnik *F_1-score* jest średnią harmoniczną precyzji i czułości – wzór (21). Gdy liczba próbek fałszywie pozytywnych oraz fałszywie negatywny jest równa 0 to *F_1-score* przyjmuje wartość równą 1 [163,165,166].

$$F_1\ score = \frac{2TP}{2TP + FP + FN} = \frac{2PPV \cdot TPR}{PPV + TPR} \tag{21}$$

Współczynnik ten łączy w sobie precyzję i czułość co czyni go bardzo użytecznym narzędziem do oceny modeli klasyfikacji binarnej. Znajduje zastosowanie w przypadkach, gdy rozkłady klas są niezrównoważone. Jednak jest on wrażliwy na zmianę oznaczania etykiet z pozytywnych na negatywne (i odwrotnie) oraz nie uwzględnia wyników prawdziwie negatywnych [163,165,166].

4.4.8. Krzywa ROC

Krzywa ROC (ang. *receiver operating characteristic*, ROC) w sposób graficzny ilustruje jakość klasyfikacji binarnej. Krzywa rysowana jest w prostokątnym układzie współrzędnych (Rysunek 4.9). No osi pionowej (*y*, rzędnych) umieszczona jest czułość (TPR), zaś na osi poziomej (*x*, odciętych) częstość fałszywych alarmów (FPR) [163–165].



Rysunek 4.9. Przykład krzywej ROC (linia żółta) wraz z zaznaczonymi ważnymi punktami oraz obszarami oznaczającymi skuteczność algorytmu: lepszą (zielony) lub gorszą (czerwony) od losowego oznaczenia (przerywana linia czerwona) [163].

Na Rysunku 4.9 zaznaczone są zaznaczone są wszystkie narożnik układu współrzędnych. Każdy punkt oznacza klasyfikator, który:

- ***A**:* nie przypisał prawidłowo żadnej etykiety pozytywnej (TPR = 0), za to cała klasa negatywna została przyporządkowana poprawnie (FPR = 0),
- ***B***: nieomylnie ocenił wszystkie próbki, pozytywne i negatywne (TPR = 1, FPR = 0),
- ***C***: oznaczył całą klasę pozytywną poprawnie (TPR = 1), natomiast, żadna z próbek negatywnych nie została przypisana prawidłowo (FPR = 1),
- ***D***: przypisał błędną klasę do każdej z próbek (TPR = 0, FPR = 1).

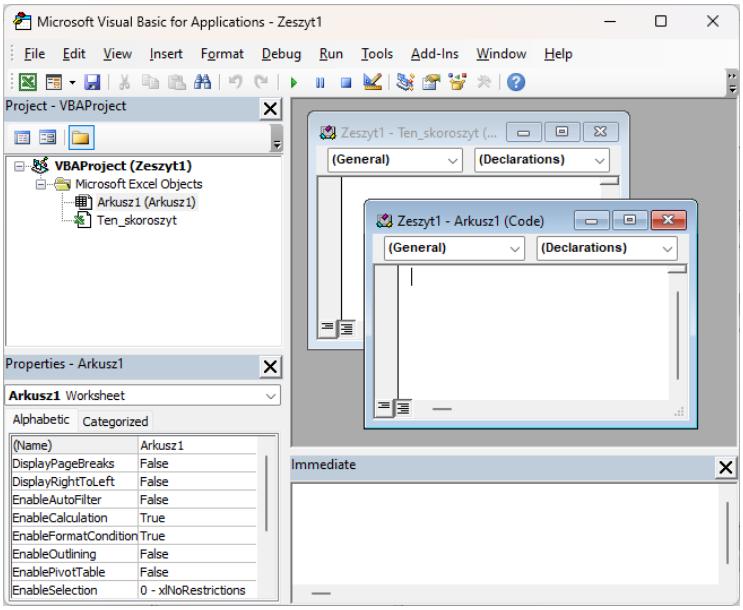
Żółta linia (Rysunek 4.9) oznacza krzywą ROC dla przykładowego klasyfikatora. Pełna krzywa jest tworzona dla algorytmu, który zawraca wartości od 0 do 1, np. klasyfikator regresji logistycznej (opisany na stronie 30). Aby stworzyć krzywą, zmieniany jest poziom (tzw. próg) od którego próbka jest klasyfikowana jako pozytywna. Naiwny klasyfikatora Bayesa nie zwraca wartości rzeczywistej, zatem nie jest też możliwe utworzenie krzywej. Natomiast sposób wizualizacji może być wykorzystany w celach prezentacji TPR i FPR [163–165].

4.5. Pakiet Microsoft Office oraz Visual Basic for Application

Firma Microsoft jest producentem powszechnie znanego oprogramowania biurowego Microsoft Office (w 2022 roku została zmieniona nazwa na „Microsoft 365”). W jego skład wchodzą takie aplikacje jak Word, Excel, PowerPoint, Outlook oraz mniej znany Access (dostępny w szerszych opcjach zakupu lub wersji skierowanej do firm). Produkt ten jest kierowany do organizacji instytucjonalnych takich jak firmy, urzędy oraz jednostki edukacyjne. W swoim raporcie rocznym za rok 2019 firma wskazała, że na świecie, z płatnej wersji pakietu korzysta 180 milionów użytkowników [167]. Aplikacje te są na tyle powszechne, że widnieją w polskim programie nauczania informatyki w klasach szkoły podstawowej [168].

W większości aplikacji z pakietu Microsoft Office (Word, Excel, PowerPoint, Outlook, Access) zaimplementowany jest język programowania Visual Basic for Applications (VBA) – Rysunek 4.10. Po raz pierwszy został on udostępniony w 1993 roku wraz z wersją MS Excel 5.0. Obecnie ciągle jest wspierany w nowych wersjach oprogramowania. VBA jest to narzędzie, które w znaczący sposób rozszerza funkcjonalność rozpatrywanej aplikacji. Z jego wykorzystaniem możliwe jest m.in. [169]:

– manipulowanie plikami tj. odczytywanie, tworzenie, usuwanie,
– tworzenie interfejsu użytkownika przez kreowanie okien dialogowych i komunikatów,
– komunikacja ze źródłami danych takimi jak relacyjne bazy danych typu SQL,
– walidacja informacji np. przez weryfikację danych wprowadzanych przez użytkownika,
– tworzenie funkcji matematycznych na bazie podstawowych operacji matematycznych.



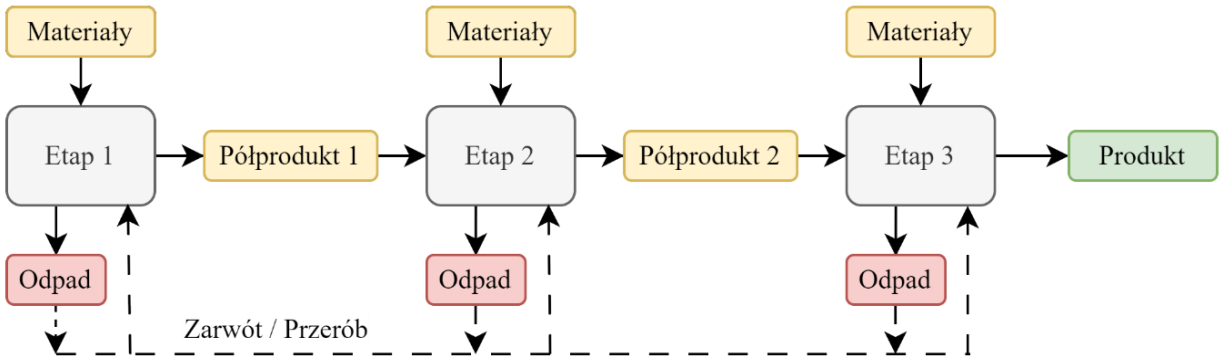
Rysunek 4.10. Okno edytora VBA otwarte w pliku Excel.

4.6. Wyroby szybkozbywalne

Produkty szybkozbywalne (ang. *Fast-Moving Consumer Good*s, FMCG) to branżowe określenie grupy produktów przeznaczonych na rynek konsumencki. Charakteryzują się one szybką sprzedażą oraz relatywnie niską ceną. Są to m.in. żywność pakowana, słodycze, środki czystości, kosmetyki, leki bez recepty (OTC) [170]. Bank Pekao S.A. raportuje, że w Polsce sektor ten pomimo pandemii oraz sytuacji geopolitycznej w 2022 roku zanotował wzrost sprzedaży rok do roku [171]. Jednak, nie bez znaczenia pozostaje fakt, że stan gospodarki w 2023 roku pozostaje niestabilny, a producenci zanotują zmniejszenie marży brutto na oferowanych przez siebie wyrobach (jednym z powodów jest presja inflacyjna) [172] – stwarza to pole do działań optymalizujących koszty produkcji.

4.6.1. Proces produkcji seryjnej w przemyśle chemicznym

Produkcja seryjna jest sposobem organizacji procesu wytwórczego, który polega na wytwarzaniu partii towarów w dokładnie ten sam sposób. Po zakończeniu cyklu produkcyjnego linie produkcyjne mogą zostać zatrzymane lub przezbrojone do produkcji innych wyrobów – w produkcji masowej wyrób jest wytwarzany w sposób ciągły z minimalizacją przestojów maszyn [173]. W trakcie trwania procesu produkcji wyrób znajduje się w ruchu pomiędzy skończoną liczbą etapów (stanowiskami produkcyjnymi), które następują po sobie kaskadowo. Produkcja seryjna charakteryzuje się: krótkimi seriami produkcyjnymi (nawet kilkugodzinnymi) oraz parkiem maszynowym projektowanym tak, aby uzyskać elastyczność produkcji (możliwość wytwarzania różnych wyrobów) [173,174]. Rysunek 4.11 ilustruje przykładowy seryjny proces produkcyjny składający się z trzech etapów:



Rysunek 4.11. Przykład schematu procesu w modelu seryjnym z trzema etapami.

Przedstawiony modelowy proces produkcji seryjnej składa się z trzech etapów. Schemat ten ilustruje ideę procesu – rzeczywisty diagram procesu produkcyjnego danego wyrobu będzie zależał od zakładu produkcyjnego, w którym jest on realizowany [175–177]. Proces rozpoczyna się od Etapu 1, który jest zasilany materiałami (surowcami), jego wynikiem jest półprodukt. Etap 2 oraz Etap 3 są stadiami pośrednim, które muszą być zasilane półproduktami. W prezentowanym modelu etapy pośrednie również wymagają dostarczania materiałów. Każdy z etapów produkcji generuje odpad, który może być produktem ubocznym lub wyrobem odrzuconym przez kontrolę jakości. Możliwe jest także zawracanie odpadów do ponownego przerobienia w celu zminimalizowania strat produkcyjnych.

Zarządzanie wyrobem niezgodnym (odpadem) ma miejsce właściwie w każdym procesie produkcyjnym. Generowanie odpadu może być spowodowane wieloma czynnikami [178]: niespełnienie kryteriów jakościowych, produkt uboczny procesu, awaria maszyny, błąd operatora itp. Odpad może zostać zutylizowany, sprzedany po zredukowanej cenie lub ponownie przerobiony. Utylizacja następuje w przypadkach, w których jakość odpadu nie jest wystarczająca by go zakwalifikować do ponownego procesowania oraz gdy koszt przerobu przewyższa koszt odpadu (uwzględniając koszt utylizacji). Sprzedaż odpadów posiadających pewną wartość: np. skrawki plastiku, uszkodzone opakowania wyrobu gotowego, wiąże się z dodatkowymi rejestracjami produktów oraz z pozyskaniem odbiorcy. Przerób w jednostce produkcyjnej („na miejscu") jest więc działaniem obecnym w prawie każdej fabryce [179–181]. Wielkość wolumenów produkcji na rynek konsumencki (np. nowa fabryka przekąsek koncernu PepsiCo w Polsce będzie przetwarzać rocznie około 10 tysięcy ton kukurydzy i 60 tysięcy ton ziemniaków [182]; fabryka farmaceutyczna koncernu GSK w Poznaniu produkuje rocznie ponad 3 miliardy tabletek oraz ponad 470 milionów kapsułek [183]) sprawia, że w procesach produkcyjnych odpad jest znaczącym kosztem – występuje efekt skali.

Nie bez znaczenie pozostaje fakt istotnego wzrostu znaczenia podejścia Zero-Odpadów (ang. *Zero-Waste*) – czyli sposobu gospodarowania odpadami, tak by wytwarzać ich jak najmniej oraz dążenie do gospodarki o obiegu zamkniętym. W krajach rozwiniętych polityka zrównoważonego rozwoju i odpowiedzialności za środowisko naturalne kreuje decyzje konsumentów, którzy coraz częściej wybierają produkty przyjazne środowisku, takiej jak te wytworzone z materiałów pochodzące z recyklingu, przechowywane w papierowych oraz szklanych opakowania, dające się recyklingować itp. A więc zarządzanie odpadem w jednostkach produkcyjnych nie jest jedynie działaniem mającym poprawić rentowność produkcji, ale również działaniem skierowanym do konsumentów by utrzymać/zdobyć udział w rynku [184–188].

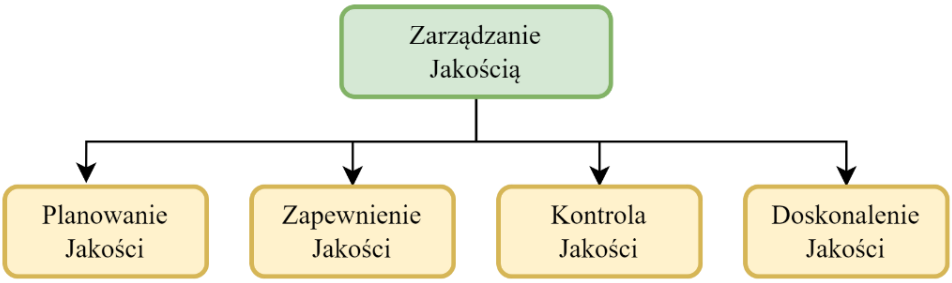
Przykładowy opis procesu produkcji realizowanego w koncernie Reckiit Benckiser

Koncern Reckitt Benckiser w większości fabryk na świecie realizuje produkcję środków czystości w dwuetapowym procesie produkcyjnym. Pierwszy etap to wytwarzanie półproduktu w modelu szarżowym. Odbywa się on w wydzielonych obszarach fabryki tzw. „obszary mieszania". Drugim etapem jest butelkowanie półproduktów na liniach pakujących.

Produkcja środków piorących polega na rozpoczęciu dozowania surowców do mieszalnika od wody sterylnej. Później dodawane są konserwanty, a następnie pozostałe surowce. Mieszalnik umieszczony jest na tensometrach, dzięki czemu automatycznie monitorowana jest wielkość dozy poszczególnych składników oraz waga całego wsadu. Surowce mogą być dozowane bezpośrednio ze zbiorników stokażowych (np. anionowe środki powierzchniowo czynne), dodawane poprzez właz umieszczony na górze mieszalnika (np. chlorek sodu) lub dozowane z wykorzystaniem armatury (np. emulgatory poprzez inżektor w ciągu cyrkulacji). Po zakończeniu dozowania wszystkich surowców zgodnie z recepturą, wykonywana jest kontrola parametrów fizykochemicznych. Następnie możliwe jest dodanie do wsadu mieszalnika pozostałości półproduktu po poprzednich wejściach produkcyjnych. Procedury koncernu wymagają dodatkowej kontroli parametrów jakościowych jeżeli operacja zawracania miała miejsce. Nie jest dopuszczane zawracanie półproduktu w innych etapach produkcji. Jeżeli kontrola jakości wykaże zgodność parametrów ze specyfikacją to wsad mieszalnika przepompowywany jest do zbiornika buforowego. W ciągu linii pakującej (butelkującej) wykonywany jest szereg operacji. Instalacja zasilana jest materiałami takimi jak butelki, korki, etykiety, kartony zbiorcze, półprodukt. W kolejności procesowej wykonywane są następujące czynności: napełnianie butelki półproduktem, aplikacja nakrętki, naklejanie etykiet (jednocześnie z obu stron), nadrukowywanie numeru partii, pakowanie butelek do kartonów zbiorczych.

4.6.2. Zarządzanie jakością w zakładzie produkcyjnym

Według międzynarodowej normy ISO 9000:2015 ogółem jakość jest definiowana jako „zdolność do zadowolenia klientów oraz przez zmierzone i niezamierzone oddziaływanie na istotne strony zainteresowane". Rysunek 4.12 przedstawia strukturę działań związanych z zarządzaniem jakością, która została zawarta w normie [189].



Rysunek 4.12. Działania związane z zarządzaniem jakością [189].

Zarządzanie jakością zostało podzielone na cztery główne obszary, których celem jest objęcie swoim działaniem wszystkich aspektów procesów od fazy projektu, przez realizację procesu, aż po wprowadzanie zmian [9,189–191]:

- **Planowanie Jakości** jest to zbiór określonych działań, celów, procesów oraz zasobów, które służą do osiągnięcia celów jakościowych. W procesach produkcyjnych może to być ustalenie wskaźnika „dobre za pierwszym razem" (DZPR), który przedstawia stosunek ilości produktu, który spełnił kryteria jakościowe w pierwszej kontroli do całości produkcji – wzór (22) [192].

$$DZPR = \frac{\left(\begin{matrix} Dobra \\ produkcja \end{matrix}\right)}{\left(\begin{matrix} Dobra \\ produkcja \end{matrix}\right) + \left(\begin{matrix} \text{Produkcja odrzucona} \\ \text{w pierwszej kontroli} \end{matrix}\right)} \tag{22}$$

- **Zapewnienie Jakości** jest to zbiór działań, których celem jest zapewnienie zaufania, że wymagania jakościowe będą spełnione, jest to np. walidacja procedur jakościowych i czynności wykonywanych w czasie produkcji, oceny ryzyka zmian.
- **Kontrola Jakości** (w polskiej wersji językowej normy, używany jest termin „Sterowanie Jakością") są to działania mające na celu zapewnienie, że kryteria jakościowe są spełnione. Takie działanie to np. kontrola poziomu napełnienia butelki przez osobę przeszkoloną na linii produkcyjnej.
- **Doskonalenie Jakości** działanie skierowanie na rozwój możliwości spełnienia kryteriów jakościowych stawianych przez interesariuszy, np. analizowanie i ulepszanie istniejących procedur. W przypadku niezgodności, analiza przyczyny wystąpienia i jej niwelacja.

Wprowadzanie zmian w procesie produkcyjnym (np. modyfikacja harmonogramu kontroli jakościowej) musi zostać ocenione przez odpowiedni zespół. Każda modyfikacja może mieć wpływ na konsumenta i jego bezpieczeństwo (np. zanieczyszczenie produktu) lub kwestie regulacyjne (np. poziom napełnienia niezgodny z wymaganiami ustawodawcy). Istnieje wiele różnych metod, które mogą służyć do ustrukturyzowanej oceny analizowanej zmiany.

Powszechnie stosowane są dwie metody, które wspierają zespoły produkcyjne oraz zarządzania jakością – analiza ryzyka FMEA (ang. *Failure Mode Effective Analysis* / pol. analiza przyczyn i skutków niepowodzenia) oraz cykl ciągłego doskonalenia PDCA (ang. *Plan, Do, Check, Act* / pol. Zaplanuj, Wykonaj, Sprawdź, Popraw) [190].

Ocena ryzyka w procesie produkcyjnym metodą FMEA

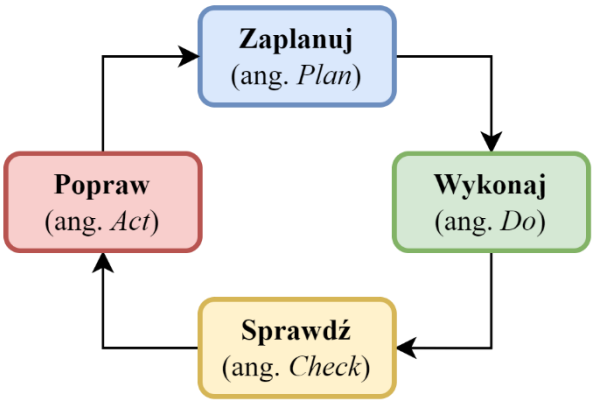
Ocena ryzyka metodą FMEA polega na definiowaniu listy wad jakie mogą wystąpić (ogólniej problemów). Przyporządkowaniu do każdej wady przewidywanych skutków jej materializacji oraz potencjalnych przyczyn jej wystąpienia. Każdą wymienioną pozycję ocenia się pod względem prawdopodobieństwa wystąpienia (niskie, średnie, wysokie), możliwości wykrycia (łatwa, średnia, trudna) oraz dotkliwości konsekwencji (niska, średnia, krytyczna). Dzięki takiej ocenie możliwe jest łatwe zdefiniowanie, które pozycje wymagają działań niwelujących (najwyższy priorytet: wysokie prawdopodobieństwo wystąpienia, trudna wykrywalność oraz krytyczne konsekwencje). Tak wykonana analiza ryzyka pozwala na znaczną redukcję prawdopodobieństwa wystąpienia niezgodności jeszcze na etapie projektowania zmiany [193–196]. Zmniejszanie ryzyka jest realizowane poprzez działania zapobiegawcze, czyli takie, których celem jest wdrożenie rozwiązań minimalizujących prawdopodobieństwo wystąpienia wady – działanie na przyczynę. Zdarzają się jednak sytuacje, w których nie jest możliwe wyeliminowanie przyczyn powstawania wady, określane są wtedy działania kontrolne, których celem jest monitorowanie czy dana wada występuje oraz czy zaistniały możliwości realizacji działań zapobiegawczych [194,197].

Jedną z głównych zalet tej metody jest to, że może ona być aplikowana właściwie do wszystkich elementów jakie składają się na proces produkcyjny: modyfikacji maszyn, zmian w procedurach, szkolenia operatorów, modyfikacji produktów, planowania produkcji itd. [196,197] W przemyśle powszechna jest konieczność posiadanie certyfikowanego systemu zarządzania jakością zgodnego z grupą norm ISO 9001, a jednym z wymagań tej normy jest ocena ryzyk i szans [10,198,199]. Klasycznie metoda FMEA przybiera formę tabeli, w której każdy wiersz to kolejna potencjalna wada. Obecnie dostępnych jest wiele rozwiązań informatycznych, które usprawniają zarządzanie ryzykiem oraz dokumentację działań.

Cykl ciągłego doskonalenia PDCA

Proces produkcyjny ulega ciągłym zmianom, które są wyzwalane przez m.in. niwelację ryzyka, wdrożenia przed procesem certyfikacji, usprawnienia z raportu audytowego, usprawniania po awariach, inwestycje w park maszynowy [1,200–202]. Zmiany w procesie

produkcyjnym również są konsekwencją braku stabilizacji, która ma miejsce we współczesnym świecie, a określana jest jako VUCA (ang. *Volatility, Uncertainty, Complexity and Ambiguity* / pol. Zmienność, Niepewność, Złożoność, Niejednoznaczność) [3,203,204]. Działania te często wymagają cyklu prób i błędów, aby potwierdzić, że przyniosły one oczekiwany efekt. Metodą powszechnie używaną w tym celu jest model ciągłego doskonalenia PDCA (Rysunek 4.13) [205,206].



Rysunek 4.13. Cykl ciągłego doskonalenia PDCA (cykl Deminga) [205].

Jest to iteracyjny model działania służący do rozwiązywania problemów oraz doskonalenia istniejących rozwiązań. Technika ta zapewnia logiczną strukturę wprowadzania modyfikacji oraz monitorowania efektów. Składa się ona z czterech etapów. Pierwszy z nich to etap „Zaplanuj" (ang. *Plan*), w którym opracowywana jest zmiana oraz sposób jej wprowadzenia, niezbędna jest również ocena aktualnej sytuacji oraz potencjalnych skutków wdrożenia. Na etapie drugim „Wykonaj" (ang. *Do*) implementowana jest koncepcja zgodnie z założeniami oraz planem stworzonym we wcześniejszym etapie. Następnie wykonywana jest weryfikacja rezultatów wprowadzonego rozwiązania – „Sprawdź" (ang. *Check*). Jeżeli zmiana przyniosła pożądane efekty to na ostatnim – czwartym – etapie „Popraw" (ang. *Act*) wykonywana jest standaryzacja rozwiązania. W przeciwnym wypadku rozpoczynany jest kolejny cykl rozpoczynający się od zaplanowania poprawek. Celem tego modelu jest osiągnięcie lepszych rezultatów niż stan obecny, a logiczne następujące po sobie etapy, strukturyzują pracę i ukierunkowują ją na osiągnięcie wyników. Duża elastyczność modelu sprawia, że jest on szeroko wykorzystywany w dziedzinach takich jak przemysł, ochrona zdrowia, lotnictwo, programowanie, budownictwo, bezpieczeństwo i higiena pracy [206–213].

Kontrola jakości w procesie produkcyjnym

Nieodłącznym elementem procesu zarządzania jakością jest Kontrola Jakości. Są to wszelkie działania, których celem jest potwierdzeni, że produkt (np. przekąska, środek

czystości, usługa) spełnia określone wcześniej kryteria. W ramach kontroli jakości przeprowadzane są testy (np. badania laboratoryjne) oraz procedury weryfikacyjne (np. inspekcje wizualne przez specjalistę ds. jakości) [10,189,199]. Istotnym aspektem jest ustalenie harmonogramu kontroli, który zawiera metody i częstotliwość weryfikacji oraz etapy procesu, w których kontrola powinna być wykonana. Powszechne jest ustalanie punktów kontroli dla: materiałów dostarczanych do zakładu, po każdym etapie produkcji oraz wyrobu gotowego. Szczegółowy plan kontroli jest ściśle zależny od typu produktu oraz od realizowanego procesu produkcyjnego. Techniki kontroli jakości przybierają rozmaite formy takie jak: wagi przepływowe w ciągu linii produkcyjnej (weryfikacja napełnienia), skanowanie kodów kreskowych/QR (weryfikacja użytych komponentów), badania parametrów fizykochemicznych, kontrola wizualna oraz sensoryczna (weryfikacja uszkodzeń opakowania, kolor, zapach, konsystencja) itp. [12,15–17,214,215]

Wykonywanie kontroli jakości jest istotne nie tylko ze względów wewnętrznych producenta, ale również z przyczyn regulacyjnych. Ustawodawstwo wielu krajów wprowadza określone normy, które wyrób musi spełniać, aby mógł być on dopuszczony do obrotu. Jedną z najbardziej podstawowych regulacji jest dyrektywa Rady Unii Europejskiej nr 76/211/EWG z dnia 20 stycznia 1976 roku [216], która dla towarów paczkowanych określa dopuszczalną tolerancję względem deklarowanej wartości napełnienia. Dyrektywa ta nakłada również na producentów obowiązek kontrolowania każdej partii produktu wprowadzanego do obrotu. W polskim systemie prawnym regulacje te obowiązują na mocy ustawy z dnia 30 września 2022 roku [217]. Istotnym aktem prawny obowiązującym na terenie Europejskiego Obszaru Gospodarczego (EOG) jest Rozporządzanie Parlamentu Europejskiego nr 528/2012 z dnia 22 maja 2012 roku, które ściśle reguluje rynek produktów biobójczych – obowiązek rejestracji, ramy kontroli w trakcie produkcji środków chemiczny do użytku konsumenckiego oraz ich obrotu na terenie EOG [11], wprowadzony w Polsce przez Sejm ustawą z dnia 2 grudnia 2020 roku [218]. Tak więc działania optymalizacyjne prowadzone w przemyśle chemicznym muszą uwzględniać szeroki kontekst funkcjonowania przedsiębiorstw, a nie jedynie efektywność ekonomiczną lub wydajność instalacji wytwórczej.

4.7. Podsumowanie części literaturowej

Przemysł dąży ku efektywności, minimalizacji ryzyka oraz poprawy jakości produktów. W ramach polityki ciągłego doskonalenia (PDCA) rewidowane są funkcjonujące narzędzia i procesy w celu ich optymalizacji oraz dostosowania do nowych potrzeb. Inżynierowie chemicy pracujący w zakładach wytwórczych również dążą do udoskonalenia technik fizykochemicznych. Prace te obejmują odpowiedni dobór metod analitycznych, częstotliwości wykonywania kontroli oraz ograniczenia kosztochłonności laboratoriów. Wymagania przemysłu oraz dostępna technologia sprawiają, że specjaliści coraz częściej wykorzystują narzędzia i metody interdyscyplinarne.

Nadzorowane uczenie maszynowe jest jedną z domen sztucznej inteligencji, w ramach której istniej bardzo wiele podejść, rozwiązań oraz węższych zakresów problemowych. Podejścia te różnią się nie tylko ogólną ideą, ale również aparatem matematycznym, korekcjami, parametryzacjami itd. To wszystko sprawia, że nawet w tak młodej dziedzinie nauki, jaką jest sztuczna inteligencja istnieje bardzo wiele rozwiązań i podejść, które mogą być adaptowane i modyfikowane na potrzeby różnorodnych badań.

Algorytmy sztucznej inteligencji mogą oceniać eksperymenty i analizy znacznie szybciej niż ludzie, charakteryzują się większą precyzją i powtarzalnością wyników. Wykorzystanie ich wpływa na poprawę bezpieczeństwa pracowników laboratoriów poprzez zawężanie zakresu eksperymentów, czyli ograniczanie czasu ekspozycji specjalistów chemików na substancje szkodliwe np. rozpuszczalniki organiczne lub wskaźniki stosowane w miareczkowaniach.

W ogólnym ujęciu problem badawczy niniejszej rozprawy ma na celu wykazać, że jest możliwa cyfryzacja jednego z elementów procesu kontroli jakości w przemyśle chemicznym w zakresie badania parametrów fizykochemicznych. Postawiona teza zakłada weryfikację działań laboratorium kontroli jakości w zakładzie produkcyjnym Reckitt Benckiser oraz wykorzystanie badań interdyscyplinarnych, które wskazują na wysokie prawdopodobieństwo:

- ograniczenia używania szkodliwych substancji chemicznych,
- przyśpieszenie klasyfikacji próbek,
- redukcji kosztów laboratorium fizykochemicznego.

Komplementarnym założeniem jest wykorzystanie standardowego komputera osobistego, udostępnianego przez przedsiębiorstwo (bez modyfikacji sprzętowej) oraz dostępnego na nim oprogramowania (bez zakupu nowych aplikacji i licencji).

Oprogramowanie, które zostało wybrane do realizacji celu badawczego to zestaw Microsoft Office. Algorytm zostanie zaprogramowany z wykorzystaniem wbudowanego

w pakiet języka programowania VBA (opisany na stronie 46). Aplikacja, która posłuży do zbudowania interfejsu użytkownika końcowego to Access, ponieważ posiada ona zaawansowany kreator okien programu (dialogowych). Ze względu na fakt, że możliwości edytora VBA są bardzo zbliżone pomiędzy poszczególnymi aplikacjami, program MS Excel również mógłby zostać wykorzystany w tym celu.

Kontrola jakości ma na celu sprawdzenie czy poddawany ocenie element jest zgodny, czy też niezgodny ze specyfikacją. Wynika z tego, że najodpowiedniejszą metodą uczenia maszynowego do realizacji celu, będzie klasyfikator binarny. Ze względu na brak domyślnych funkcji w języku VBA, szczególnie trudne i mozolne było tworzenie klasyfikatora na bazie algorytmu drzewa klasyfikacyjnego (opisany na stronie 32) oraz losowego lasu decyzyjnego (opisany na stronie 33). Nie mogą również zostać zastosowane algorytmy wykorzystujące operacje matematyczne wymagające znacznej mocy obliczeniowej. Takie obliczania są wykorzystywane w algorytmie regresji logistycznej (funkcja logitowa, logarytmowanie; opisany na stronie 30) oraz w modelu *k*-najbliższych sąsiadów (potęgowanie, pierwiastkowanie; opisany na stronie 31). Tak więc najodpowiedniejszy typ algorytmu klasyfikacji binarnej do realizacji celu badawczego to naiwny klasyfikator Bayesa (opisany na stronie 34). Należy także zaznaczyć, że ten typ klasyfikacji znajduje szerokie zastosowanie i uznanie.

5. Część eksperymentalna

Przedmiotem niniejszej rozprawy doktorskiej jest wykorzystanie narzędzi statystycznych w analizach jakościowych, w obszarach produkcyjnych. Praca miała na celu zbadanie możliwości stworzenia narzędzia z dziedziny uczenia maszynowego. Na drodze do realizacji pracy sprawdzono wymagania zakładu produkcyjnego oraz wybrano etap procesu kontroli jakości, którego automatyzacja była obarczona najmniejszym ryzykiem. Na bazie przeglądu literaturowego oraz wymagań przedsiębiorcy sformułowane zostały założenia części eksperymentalnej oraz koncepcja działania algorytmu. Niniejsza część zawiera omówienie przyjętych założeń, opis algorytmu klasyfikacji, przedstawienie funkcjonalności aplikacji oraz metody oceny działania ostatecznego narzędzia.

5.1. Założenia do części eksperymentalnej

Rozprawa skupia się na opracowaniu oraz zweryfikowaniu działania narzędzia statystycznego, którego celem jest automatyzacja jednego z etapów procesu kontroli jakości. Precyzując, oceny parametrów fizykochemicznych półproduktu, który może zostać ponownie wykorzystany w procesie produkcyjnym.

Sieci handlowe i dystrybucyjne składają zamówienia na produkty w liczbie sztuk (liczba butelek). Fabryka realizuje proces produkcyjny w taki sposób, aby dostarczyć zamówiony towar w dokładnie żądanej ilości. Mniejsza liczba dostarczonych opakowań konsumenckich niż zamówiona wiąże się z karami za brak realizacji, zaś większa z kosztami magazynowania oraz ryzykiem starzeniem się produktu. Elementy opakowaniowe nie starzeją się w znaczącym tempie oraz możliwe jest ich tanie magazynowanie na terenie fabryki. Powszechne jest również używanie takich samych materiałów (np. butelek, korków, kartonów) do wielu wyrobów oraz marek. Natomiast nie jest możliwe dokładne zsynchronizowanie procesu wytwarzanie półproduktów płynnych z zamówioną ilości butelek. Wiąże się to z minimalną wielkością wsadu mieszalnika oraz startach na transferze pomiędzy instalacjami. Przez to powszechną praktyką jest wytwarzanie większej ilości półproduktu, niż to wynika z zamówienia. W ramach

optymalizacji kosztów oraz redukcji marnotrawstwa półprodukty, które zastały w zbiornikach magazynowych po zrealizowanym procesie butelkowania są ponownie wykorzystywane. Według procedur koncernu wymagają one jednak weryfikacji parametrów fizykochemicznych przed ponownym użyciem. Klasyfikacja serii półproduktu do ponownego wykorzystania w procesie produkcyjnym w zakładzie Reckitt Benckiser Production (Poland) wykonywana jest przez inżyniera chemika.

Pracę podzielono na trzy etapy badawcze. Celem pierwszego z nich było stworzenie koncepcji algorytmu klasyfikacji binarnej, czyli szczegółowej definicji kluczowych aspektów pracy badawczej, w postaci:

- oceny dostępności oraz jakości danych produkcyjnych,
- analizy ryzyk dotyczących implementacji narzędzia oraz definicja działań niwelujących,
- podstawowych funkcjonalności narzędzia,
- zakresu funkcjonalności interfejsu użytkownika,
- kryterium sukcesu w kontekście celu pracy oraz potrzeb przedsiębiorcy.

Realizację tej fazy badań przeprowadzono poprzez analizę rozpatrywanego procesu produkcyjnego oraz towarzyszących mu czynności zarządzania jakością. Efekty tego etapu przedstawione zostały w Rozdziale 5.2.

Drugi etap badawczy obejmował stworzenie algorytmu wraz z interfejsem użytkownika według opracowanej w pierwszej fazie badań koncepcji. Jako środowisko programowania została wybrana aplikacja Access (Microsoft Office), aby zrealizować cel użycia tylko i włącznie standardowo dostępnych zasobów. Kluczową funkcjonalnością, która zaważyła na wyborze tego programu były obszerne możliwości tworzenia formularzy (okien dialogowych). Wykorzystano algorytm naiwnego klasyfikatora Bayesa. Został on zaprogramowany w języku VBA, a jego parametry sterujące zostały dobrane na bazie dostępnych danych produkcyjnych. Opracowana aplikacja koncepcyjnie zastępuje element kontroli jakości, dlatego konieczne było stworzenie graficznego interfejsu użytkownika, by umożliwić personelowi produkcyjnemu obsługę nowego narzędzia. Ponieważ wytwarzane produkty podlegają częstym zmianom, niezbędne było zaimplementowanie możliwości sterowania algorytmem. Efektem prac tego etapu było kompletne narzędzie – algorytm klasyfikacji wraz z interfejsem użytkownika opisany w Rozdziale 5.3.

Ostatni etap prac badawczych zakładał zdefiniowanie optymalnych parametrów algorytmu oraz określenie czy spełnione zostały kryteria sukcesu. W tym celu stworzony został

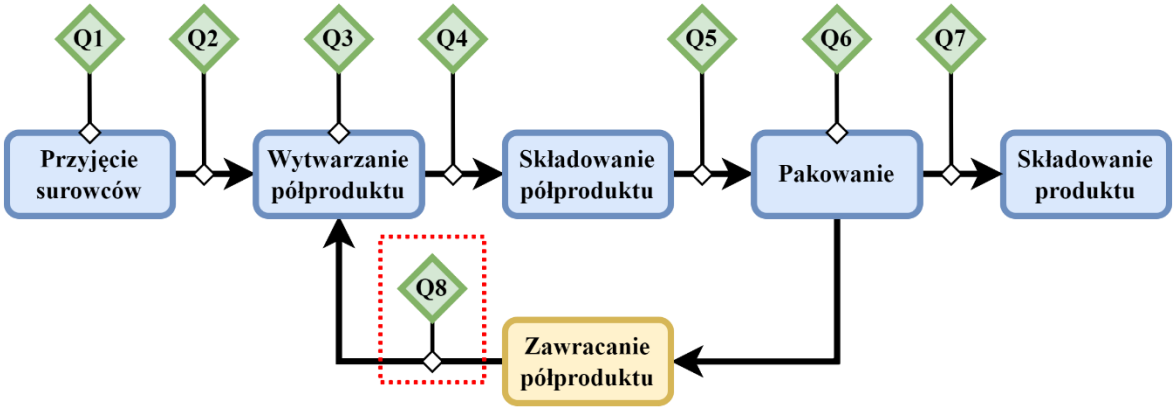
plan testów (Rozdział 5.4), który zwierał wybrane wskaźniki oceniające jakość klasyfikacji. Został ustalony punkt odniesienia w postaci arbitralnie dobranych atrybutów sterujących. Następnie wszystkie parametry kolejno podlegały iteracyjnym zmianom wartości względem punktu odniesienia. Po każdej modyfikacji uruchamiana była ewaluacja, aby uzyskać wartości wskaźników oceniających algorytm. W ten sposób otrzymano charakterystykę wpływu danego parametru na jakość klasyfikacji. Następnie wysterowane zostały wszystkie atrybuty jednocześnie uwzględniając optymalne wartości z poprzednich testów.

5.2. Koncepcja algorytmu oraz aplikacji komputerowej

W tej części pracy przedstawiono analizę rozpatrywanego procesu produkcyjnego oraz towarzyszących mu działań jakościowych. Na jej podstawie określono, który punk kontrolny może zostać poddany automatyzacji z wykorzystaniem algorytmu uczenia maszynowego. Scharakteryzowano dostępne dane produkcyjne – zbiór treningowy. W koncepcji zawarto również wymagania względem oprogramowania wynikające z działalności prowadzonej przez zakład produkcyjny. Tę fazę badań zakończono określeniem kryteriów sukcesu.

5.2.1. Dobór elementu procesu kontroli jakości

Podstawą do stworzenia koncepcji narzędzia statystycznego była analiza procesu produkcyjnego. W tym celu stworzono diagram reprezentujący przepływ materiałów, etapy wytwórcze oraz punkty kontroli jakościowej w zakładzie Reckitt Benckiser (Rysunek 5.1).



Rysunek 5.1. Diagram analizowanego procesu produkcyjnego z punktami kontroli jakościowej (Q) oraz ze wskazanym elementem będącym przedmiotem badań (czerwona ramka przerywaną linią).

5.2.2. Opis procesu produkcyjnego

Przyjęcie surowców

Surowce do produkcji płynnych środków czystości były przechowywane na terenie fabryki w trzech rodzajach magazynów. Pierwszym z nich był stokaż, który składał się ze zbiorników do przechowywania cieczy o pojemności od 15 m^3 do 50 m^3. Dostawy odbywały się z wykorzystanie cystern samochodowych. Magazynowane tam były m.in. takie surowce jak roztwór podchlorynu sodu (wykorzystywany w środkach czystości stosowanych w łazienkach), roztwór nadtlenku wodoru (wykorzystywany w wybielających środkach czystości), prekursory surfaktantów anionowych (np. roztwór kwasu alkilobenzenosulfonowego, ABS), wodę i inne.

Drugim typem magazynu była konstrukcja służąca do przechowywania kompozycji zapachowych, które charakteryzowały się niską temperaturą zapłonu, a więc stwarzały ryzyko pożarowe. Przechowywane były więc w beczkach stalowych oraz na otwartej, zadaszonej konstrukcji stalowej, która w przypadku zapłonu kierowała siłę wybuch w określonym normami bezpieczeństwa kierunku.

Ostatnim rodzajem magazynu był standardowy budynek magazynowy, w którym przechowywano materiały dostarczane na paletach. Były to surowce sypkie (np. worki lub tzw. big-bagi) oraz surowce płynne (np. beczki lub paletopojemniki).

Wytwarzanie półproduktu

Fabryka wytwarzała półprodukty płynne w modelu wsadowym (tzw. szarżowym). W tym celu wykorzystywane były mieszalnik o pojemności od 5 m^3 do 25 m^3 (zakład dysponował ponad 20 takimi instalacjami). Każdy z nich był wyposażony mieszadło umieszczone centralnie oraz system cyrkulacji wsadu (pompa pobierała płynu od dołu mieszalnika i przepompowywała go na górę). W ciągu obiegu cyrkulacyjnego umieszczony był kolektor zaworowy z podłączonymi rurociągami od zbiorników stokażowych. Odpowiednie wysterowanie zaworów pozwalało na dozowanie surowców. Aby zapewnić bezpieczeństwo instalacje były specjalizowane, to znaczy, że jeżeli istniało ryzyko niepożądanych reakcji pomiędzy surowcami (np. reakcja NaClO z H_2O_2) to nie podłączano ich do jednego mieszalnika.

Uzbrojenie w armaturę zależało od typu wytwarzanych receptur. W przypadku procesu mieszania, podczas którego zachodziła reakcja neutralizacji, która wydzielała ciepło (np. wytwarzanie anionowego związku powierzchniowo czynnego SLES z prekursora ABS), mieszalnik był wyposażony w system chłodzenia (stosowano wymienniki płytowe). Do dozowania sypkich surowców trudno rozpuszczalnych używano pomp mieszających.

Monitorowanie dozy surowca było realizowane na różne sposoby. Każdy mieszalnik był wyposażony w tensometry, a więc wielkość dozy surowca (sypkiego lub płynnego) jak i całego wsadu mieszalnika była monitorowana przez operatora. W uzasadnionych przypadkach stosowane były przepływomierze, które zapewniały większą dokładność pomiaru (np. dla konserwantów płynnych, dla których prawo określa górną granicę dopuszczalnego stężenia). Jeżeli ilość surowca była niewielka (mniejsza od 50 kg) przed operacją mieszania przygotowywane były naważki (np. barwniki, kompozycje zapachowe).

Wytwarzanie odbywało się w modelu wsadowym. Do każdej szarży operator mieszalnika otrzymywała kartę recepturową. W dokumencie były wyszczególnione wszystkie surowce wraz z dopuszczalnymi odchyłkami od ilości dozowania, kolejnością ich dodawania oraz czynnościami jakie należy wykonać. Wszystkie działania były wyszczególnione chronologicznie, mogło to być: dozowanie surowca w określonej ilości, podgrzanie wsadu do określonej temperatury, uruchomienie cyrkulacji, wykonanie pomiaru pH, przekazanie próbki do laboratorium fizykochemicznego, kontrola wizualna itp.

Wytwarzanie każdej szarży kończyło się przekazaniem próbki do laboratorium w celu kontroli parametrów fizykochemicznych. Zakres badań różnił się w zależności od rodzaju produktu. Przed otrzymaniem wyników z laboratorium operator nie mógł podjąć żadnych działań. Jeżeli wyniki oznaczania parametrów były zgodne ze specyfikacją operator mógł przesłać wsad na zbiornik magazynowy lub wykonać dodatkową operację, czyli dodanie półproduktu, który pozostał z poprzedniego butelkowania.

Ponadto instalacje produkcyjne (niezależnie od etapu) oraz obszary były dostosowane do parametrów fizykochemicznych wytwarzanych półproduktów. Tak więc:

- instalacje mające kontakt z podchlorynem sodu wykonane były z tworzywa i/lub tytanu,
- elementy stalowe mające kontakt z nadtlenkiem wodoru poddawane były pasywacji,
- obszar i instalacje wykorzystywane do produkcji formuł wrażliwych na zakażenie mikrobiologiczne były budowane i utrzymywane w podwyższonym standardzie czystości,

Składowanie półproduktu

Każdy mieszalnik połączony był z wieloma zbiornikami magazynowymi. Taki układ instalacji pozwalał na zbudowanie zapasu jednego półproduktu oraz do wytwarzania różnych półproduktów na tym samym mieszalniku. Pierwsza sytuacja była konieczna w przypadku, gdy linia pakująca miała wyższą wydajność niż odpowiadający jej mieszalnik. Miało to miejsce szczególnie w przypadku półproduktów charakteryzujących się dużą lepkością oraz

jednocześnie pakowanych w butelki o dużej pojemności (większe niż 1 litr) – były to np. płyny do prania lub do czyszczenia toalet. Takie receptury wymagały więcej czasu na homogenizację (większa lepkość oznaczała dłuższe rozprowadzanie surowca w objętości wsadu). Z drugiej strony były mniej podatne na pienienie się podczas dozowania do butelki, przez co wydajność linii była wysoka. W takiej sytuacji etap wytwarzania półproduktu musiał rozpocząć się odpowiednio wcześniej, tak by zbudować zapas, który pozwalał linii pakująca na butelkowanie całego zamówienia bez przestojów. Druga sytuacja zachodziła, gdy na dwóch (lub więcej liniach) butelkowane były różne półprodukty wytwarzane na tym samym mieszalniku (np. odświeżacze powietrza, które różniły się jedynie typem kompozycji zapachowej).

Pakowanie

Mianem pakowania określano czynności wykonywane szeregowe, które prowadziły do wytworzenia produktu gotowego. W analizowanym zakładzie do linii pakującej na paletach dostarczane były materiały opakowaniowe (butelki, nakrętki, etykiety, kartony zbiorcze). Półprodukt transportowany był ze zbiorników magazynowych do zbiorniczków buforowych za pomocą rurociągów. Każda linia pakująca podłączona była do wielu zbiorników poprzez wyspy zaworowe. Na początku ciągu technologicznego pakowania dostarczane były butelki, które pozycjonowano na transporterze taśmowym. Następnie wprowadzano je do maszyny (tzw. nalewarki), której zadaniem było napełnienie butelki do zadanej objętości. Stosowano trzy rodzaje maszyn:

- Wyposażone w cylindry z mechanicznie regulowanym tłokiem. W jednym takcie maszyna wykonywała dwie czynności, napełnia cylinder półproduktem ze zbiorniczka buforowego, a następnie opróżniała cylinder do butelki.
- Kontrolujące poziom napełnienia butelki za pomocą przepływomierza, który był zamontowany przy każdym nalewaku wprowadzanym do butelki.
- Wykorzystujące szalki wagowe. Każda butelka pozycjonowana była na wadze, a napełnianie kończyło się po uzyskaniu zadanej masy butelki wraz z półproduktem.

Butelki po napełnieniu kierowano za pomocą transporterów do maszyny, która nakręcała korek. Następnym etapem było nałożenie etykiety przedniej i tylnej. W kolejnej czynności nadrukowywano numer partii. Po tym etapie opakowanie konsumencie było kompletne. Zadaniem dalszej maszyny w ciągu technologicznym było zgrupowanie butelek i umieszczenie ich w kartonie zbiorczym, które to następnie było układane na palecie.

5.2.3. Opis procesu kontroli jakości

Przedsiębiorstwo w celu badań do niniejszej rozprawy udostępniło obszar wytwarzania środków czystości w postaci płynnej (np. wybielacze, odświeżacze powietrza, płyny do prania). Przeanalizowany został proces wytwórczy, a następnie sporządzono jego diagram (Rysunek 5.1). Zaobserwowana sekwencja rozpoczynała się od odebrania dostaw surowców. Wewnętrzne regulacje kontroli jakości obligowały partnerów biznesowych fabryki, do tego, by każda partia materiału w dokumentach transportowych zawierała świadectwo wykonanej przez nich kontroli. Pierwszą czynnością jakościową (oznaczoną jako Q1) była weryfikacja dokumentów. Jej następstwem była informacja w systemie zarządzania magazynem, czy dana partia surowca została udostępniona (zwolniona) do produkcji. Znikoma część surowców (mniej niż 10% ogólnej liczby używanych pozycji) przechodziła kontrolę w trakcie przekazywania materiału z magazynu wewnętrznego na obszar produkcji (Q2). Spośród wszystkich wyspecyfikowanych parametrów sprawdzano tylko jedną kluczową cechę (np. zawartość procentową głównego składnika). Oznaczoną wartość archiwizowano w dokumencie produkcyjnym (tzw. karcie szarży) oraz w pliku komputerowym typu Excel.

Półprodukt wytwarzano w modelu wsadowym (tzw. szarżowym). Zgodnie z recepturą (kartą szarży) do mieszalnika dozowano surowce oraz uruchamiano operacje pomocnicze (np. podgrzewanie). W trakcie tego etapu wykonywano analizy parametrów fizykochemicznych (Q3) np. pomiar lepkości. Zakres badań zależał od typu półproduktu oraz wykorzystywanej instalacji. Zapisy prowadzono w sposób nieustandaryzowany w plikach elektronicznych oraz dokumentach papierowych. Przed transferem wytworzonego półproduktu do zbiornika magazynowego wykonywano analizę wszystkich cech jakościowych (Q4), zgodnie z odpowiadającą specyfikacją. Jeżeli okres składowania był dłuższy niż jedna godzina, to przed rozpoczęciem kolejnego etapu powtarzano pełną kontrolę jakościową (Q5). Wyniki obu sprawdzeń zapisywano w pliku Excela oraz odnotowywano w dokumentach.

Podczas pakowania, czyli napełniania opakowań konsumencki (butelek) wytworzonym wcześniej półproduktem, kontrolowano niektóre parametry fizykochemiczne (Q6). Zakres tych badań nie był wystandaryzowany. Zależał on od typu wyrobu (np. rozmiaru opakowania, kraju przeznaczenia), półproduktu oraz instalacji, na której realizowano pakowanie. Archiwizację wyników tych sprawdzeń prowadzono w sposób niejednorodny. W zależności od rodzaju mierzonej wartości ocena mogła być zapisana w bazie danych SQL, pliku Excel lub dokumentacji papierowej.

Kontrolę jakości oznaczoną jako Q7 wykonywano po zakończeniu pakowania. Miała ona na celu potwierdzenie, że wytwarzano produkt spełniający warunki specyfikacji. Pozwalała ona na przekazywanie wyrobu gotowego do magazynu. Wykonywano ją ponownie, jeżeli w trakcie realizacji produkcji nastąpiła zmiana personelu. Wyniki tej kontroli przechowywano w pliku Excela oraz odnotowywano w dokumentach papierowych.

Nie występowała pełna synchronizacja etapów produkcji. Przez co ilość wytworzonego półproduktu była większa niż zapotrzebowanie procesu pakowania. Nadmiarową ilość kierowano ponownie do produkcji tego samego półproduktu w celu zminimalizowania kosztów oraz marnotrawstwa. W celu dopuszczenia do ponownego wykorzystania poprzedniej partii płynu, wymagana była kontrola jego parametrów fizykochemicznych (Q8). Specjalista oceniał wyniki analiz laboratoryjnych oraz wizualny wygląd próbki, aby zakwalifikować kontrolowaną partię. W przypadku skierowania pozostałości półproduktu do ponownego użycia, jego ilość dodana do nowej szarży nie mogła przekroczyć 10% masy produkowanego wsadu mieszalnika. Wyniki kontroli jakościowej oraz decyzję specjalisty zapisywano w bazie danych SQL.

Ocena możliwości cyfryzacji i automatyzacji punktów kontroli jakości

Analiza procesu produkcyjnego pozwoliła określić w jaki sposób realizowano kontrole jakości – punkty od Q1 do Q8 (Rysunek 5.1). Zweryfikowany został również sposób archiwizowania danych. Do opracowania narzędzia klasyfikującego, opartego na algorytmie uczenia maszynowego, wykorzystuje się dane treningowe. Taka baza informacji powinna zawierać odpowiednią liczbę próbek, przypisane do nich parametry oraz etykiety (wyniki klasyfikacji). Poszczególne kontrole jakości zostały ocenione pod względem kompletności i dostępności danych historycznych.

Sposób prowadzenia kontroli jakości oraz archiwizacji danych w punktach Q1, Q2, Q3 oraz Q6 w formie zastanej nie pozwalał na ich cyfryzację i automatyzację. W tych przypadkach nie istniały kompletne cyfrowe zbiory danych historycznych, na bazie których mógłby zostać opracowany algorytm uczenia maszynowego.

Zbudowanie algorytmu uczenia maszynowego okazało się możliwe dla punktów kontroli Q4, Q5, Q7 oraz Q8. W ramach tych czynności wykonywano wystandaryzowane analizy parametrów fizykochemicznych. Każda kontrola posiadała jednolity sposób archiwizowania danych w plikach elektronicznych. Dostępne były informacje takie jak: dane półproduktu, wartości zbadanych parametrów oraz wynik kontroli jakości. Pliki, w których przechowywano dane mogły być bezpośrednio użyte jako zbiór treningowy dla aplikacji klasyfikującej.

Punkt kontroli półproduktu zawracanego do produkcji (Q8) został wybrany jako przedmiot badań. Głównym aspektem, który zadecydował o objęciu cyfryzacją tego elementu była minimalizacja ryzyka dla zakładu współpracującego. W przypadku decyzji o wdrożeniu opracowanego narzędzia, to nieprawidłowa klasyfikacja wykonana przez algorytm niosłaby najmniejsze zagrożenie dla jakości wyrobu gotowego, ponieważ:

- zawracany półprodukt uprzednio przechodził wielokrotnie kontrolę jakości,
- dodawana partia mogła stanowić maksymalnie 10% masy wsadu mieszalnika, a więc ograniczano wpływ dodatku na parametry fizykochemiczne całej szarży,
- po zawróceniu półproduktu, a następnie zakończeniu produkcji szarży, wykonywano pełną kontrolę parametrów – możliwość wykrycia potencjalnej niezgodności.

5.2.4. Dane produkcyjne

Przeanalizowane zostały dane produkcyjne dostępne w pliku komputerowym (baza danych produkcyjnych). Zestawienie dostępnych parametrów przedstawiono w Tabeli 5.1. Wszystkie argumenty zostały scharakteryzowane pod względem tego jakie wartość mogą być im przypisane w bazie danych: tekst (dowolna wartość tekstowa lub liczbowa), prawda/fałsz, data oraz słownik (zamknięta lista wyboru, której użytkownik nie może modyfikować). Dodatkowo każdemu parametrowi przypisano typ wartości: D (zmienna dyskretna) oraz R (argument z zakresu liczb rzeczywistych).

W algorytmie wykorzystano wszystkie parametry oznaczone jako dyskretne (Tabela 5.1) w sposób bezpośredni, czyli nie poddawano ich dodatkowym operacjom. W obliczeniach prawdopodobieństwa nie uwzględniono argumentów przybierających wartości dat. Jednak zastosowano je w celu ograniczenia zakresu czasowego, z którego pobierano informacje o próbkach (ograniczenie zbioru treningowego). Parametry rzeczywiste poddawano transformacji do zmiennych dyskretnych. W produkcyjnej bazie danych dla analiz fizykochemicznych nie prowadzono weryfikacji, czy wprowadzane wartości stanowią liczby. Przykład stanowią zapisy wyników analizy lepkości, w których odnajdowane są informacje tekstowe, np. „EEE”. Powoduje to konieczność uwzględnienia w obliczeniach jedynie poprawnie wprowadzonych wartości liczbowych.

Baza danych produkcyjnych udostępniona do badań zawierała informacje gromadzone od początku 2012 roku do listopada 2022 roku. Zawierała ona dane o 538 półproduktach. Całkowita liczba próbek wyniosła 44 236. Pozycje poddawano klasyfikacji wykonywanej przez specjalistów (personel posiadający odpowiednie uprawnienia). Próbkom przypisano klasę

pozytywną (zawrócić: ponowne wykorzystanie w procesie produkcyjnym) 39 757 razy, zaś 2 618 razy negatywną (ściek: seria przeznaczona do utylizacji). Bez oceny pozostało 1861 pozycji. Klasyfikacja półproduktu nie była jednoznaczna. Stosowano również warunkowe zawracanie wyrobu. W takim przypadku do pozytywnej oceny próbki („zawrócić") dołączano obowiązkowe zalecenia np. „przefiltrować przed zawróceniem". Takie zaklasyfikowanie stanowiło 15,9% wszystkich ocen pozytywnych (7 038 z 39 757).

Tabela 5.1. Parametry dostępne w bazie produkcyjnej.

Nazwa parametru (nazewnictwo produkcyjne)	Typ danych	Typ wartości	Opis
Numer próbki	Tekst	D	Numer porządkowy identyfikujący pobraną próbkę w bazie danych
Kod Wyrobu	Tekst	D	Kod alfanumeryczny, który jednoznacznie identyfikuje półprodukt
Nazwa produktu	Tekst	D	Słowna nazwa produktu
Data pobrania	Data	R	Data, kiedy została pobrana próbka do analiz laboratoryjnych
Termin przydatności	Data	R	Data, do kiedy dany półprodukt może zostać zużyty
Mikrobiologiczna	Prawda/Fałsz	D	Informacja czy półprodukt jest wrażliwy na zakażenie mikrobiologiczne
Rodzaj	Słownik	D	Klasyfikacja półproduktu ze względu na jego pochodzenie (np. produkcja, test, mycie instalacji)
Typ	Słownik	D	Oznaczenie czy wyrób jest normatywny czy niezgodny
Instalacja	Słownik	D	Nazwa instalacji pochodzenia
Kolor	Prawda/Fałsz	D	Wynik analizy sensorycznej na zgodność ze wzorcem
Zapach	Prawda/Fałsz	D	Wynik analizy sensorycznej na zgodność ze wzorcem
Wygląd	Prawda/Fałsz	D	Wynik analizy sensorycznej na zgodność ze wzorcem
pH	Tekst	R	Wartość pomiaru pH
Gęstość	Tekst	R	Wartość pomiaru gęstości
Lepkość	Tekst	R	Wartość pomiaru lepkości
Nadtlenku wodoru	Tekst	R	Wartość stężenia procentowego nadtlenku wodoru
Chlor	Tekst	R	Wartość stężenia procentowego wolnego chloru
Sucha pozostałość	Tekst	R	Wartość suchej pozostałości, wyrażana w procentach masy początkowej
Decyzja	Słownik	R	Decyzja jaką może podjąć technolog: zawrócić/ściek
Technolog	Słownik	D	Lista osób uprawnionych do decyzji
Zalecenia	Tekst	D	Działania jakie technolog zleca, np. zmniejszenie ilości możliwej do zawrócenia

5.2.5. Wymagania stawiane aplikacji komputerowej

Zakład produkcyjny, w którym realizowane były badania, posiadał certyfikowany system zarządzania jakością (SZJ), zgodny z normą ISO 9001:2015. Standaryzacja ta wymagała regulacji wewnętrznych (procedur), modelu podejmowania decyzji oraz dokumentowania działań. W fazie koncepcyjnej zostały określone ryzyka niezgodności, a w ich następstwie kryteria, które nowe oprogramowanie jest zobowiązane spełnić, aby zostało dopuszczone do użytku. W Tabeli 5.2 zestawiono wymagania oraz sposób ich realizacji, tak aby została zachowana zgodność z procedurami – działania niwelujące ryzyko.

Tabela 5.2. Wymagania względem oprogramowania oraz przyjęty sposób ich realizacji.

Wymaganie systemu zarządzania jakością	**Sposób realizacji (funkcja aplikacji)**
Pierwotne dane produkcyjne nie mogą być zmieniane	Opracowana została nowa aplikacja, ta już istniejąca nie była modyfikowana. Połączenie z bazą danych realizowane jest w trybie „tylko odczytu”.
Decyzje jakościowe mogą być podejmować tylko przez ściśle wyznaczone osoby	Wynik algorytmu klasyfikacji nie jest zapisywany automatycznie. Wymagane jest użycie przycisku „zapisz” przez człowieka. Lista uprawnionych osób jest pobierana z produkcyjnej bazy danych.
Wprowadzone zmiany muszą być identyfikowalne	Aby zapisanie wyniku było możliwe konieczne jest wybranie z listy oznaczenia identyfikującego osobę. Data i czas zapisywane są automatycznie.
Korekcja ustawień musi być zapisywana w rejestrze	Zmiana któregokolwiek parametru sterującego algorytmem wywołuje komunikat ostrzeżenia. Jeżeli zmiana będzie kontynuowana to automatycznie wygeneruje się plik zawierający: parametry przed zmianą, wskaźniki oceny algorytmu, dane każdej ocenionej próbki.
Użytkownikowi musi być zapewniony dostęp do aktualnych informacji	W aplikacji prezentowane są pierwotne dane próbki, wyniki obliczeń algorytmu, przyznana klasa, dane użyte przez algorytm, bieżące wartości wskaźników oceny narzędzia oraz komunikaty.
Muszą istnieć możliwości analizy, monitorowania i doskonaleniu procesu	Wyświetlanie są wskaźniki oceny narzędzia i dane, na podstawie których zostały one obliczone. Po każdej zapisanej klasyfikacji są one aktualizowane. Możliwa jest zmiana parametrów algorytmu oraz uruchomienia ewaluacji z danych historycznych.

5.2.6. Wymagania stawiane algorytmowi

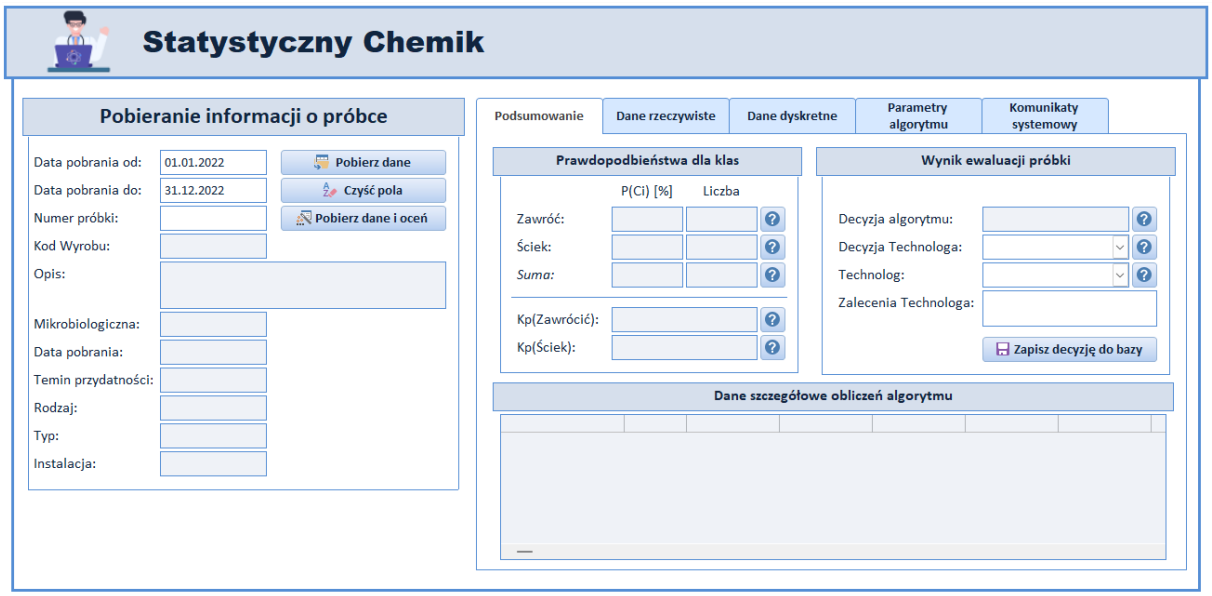
W Rozdziale 4.7 „Podsumowanie części literaturowej”, jako model algorytmu uczenia maszynowego został wskazany naiwny klasyfikator Bayesa. Zaprogramowany go w aplikacji Access (Microsoft Office) z wykorzystaniem edytora języka VBA. Dokładność klasyfikacji (ACC; opisana na stronie 42) została wskazana jako główny wskaźnik oceny algorytmu uczenia maszynowego – opracowywanego narzędzia. Ustalone zostały dwa kryteria sukcesu.

Pierwsze z nich dotyczyło, przedmiotu dociekań niniejszej rozprawy, czyli badań nad możliwościami wykorzystania narzędzi statystycznych w analizach jakościowych, w obszarach produkcyjnych. Jeżeli dokładność wyniesie powyżej 50% to odpowiedź będzie twierdząca. W takiej sytuacji poziom poprawnych predykcji będzie większy niż w przypadku losowego przypisania klasy. W badanym rozwiązaniu możliwe jest otrzymanie wyniku spośród dwóch możliwości (klasyfikator binarny). Tak więc dla zdarzeń wyłącznie losowych prawdopodobieństwo dla jednej z możliwości wynosi ½ czyli 50% (np. szansa wylosowania reszki w pojedynczym rzucie monetą).

Drugie kryterium sukcesu dotyczyło możliwości implementacji narzędzia w procesie kontroli jakościowej. Ustalono, że wartość dokładności powyżej 95% sprawi, że narzędzie będzie posiadało możliwości wdrożeniowe. Odpowiednio wysoki poziom skuteczności pozwala niwelować ryzyko dla jakości produktu oraz bezpieczeństwa konsumentów. Ustalony został również próg zdolności oznaczenia próbek (ASR) pochodzących ze zbioru treningowego (analiz historycznych) na poziomie minimum 80%.

5.3. Opis funkcji aplikacji

W tym rozdziale rozprawy przedstawiono funkcje opracowanego narzędzia. Prezentowane identyfikatory produktów i instalacji zostały zanonimizowane by zachować tajemnicę przedsiębiorstwa. Zgodnie z wymaganiami (Tabela 5.2) dane źródłowe oraz istniejącą aplikację produkcyjną zabezpieczono przed modyfikacją. W tym celu opracowano nowy, niezależny program. Na Rysunku 5.2 przedstawiono stronę startową narzędzia. Opracowano je w całości, w toku realizacji prac badawczych. Programowi nadano nazwę „Statystyczny Chemik".



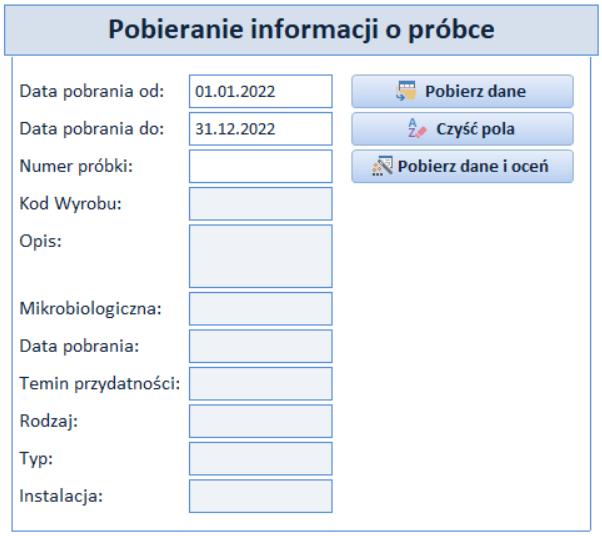
Rysunek 5.2. Opracowane narzędzie „Statystyczny Chemik" – okno początkowe.

Ekran początkowy opracowanej aplikacji przedstawiono na Rysunku 5.2 Interfejs użytkownika podzielono na dwa obszary. Pierwszy z nich znajdujący się po lewej stronie, zawiera się w ramce o nagłówku „Pobieranie informacji o próbce". Wykorzystywany jest do ściągania informacji z bazy danych oraz do ciągłego prezentowania kluczowych parametrów personelowi produkcji. Drugi obszar znajdujący się po prawej stronie, zbudowano z 5 kart:

- Podsumowanie (wyniki obliczeń pośrednich oraz klasyfikacji),
- Dane rzeczywiste (informacje pierwotne pobrane z bazy danych),
- Dane dyskretne (parametry po dyskretyzacji),
- Parametry algorytmu (sterowanie programem oraz wyniki jego ewaluacji),
- Komunikaty systemowe (informacje o zakończonych etapach i błędach).

5.3.1. Pobieranie danych produkcyjnych

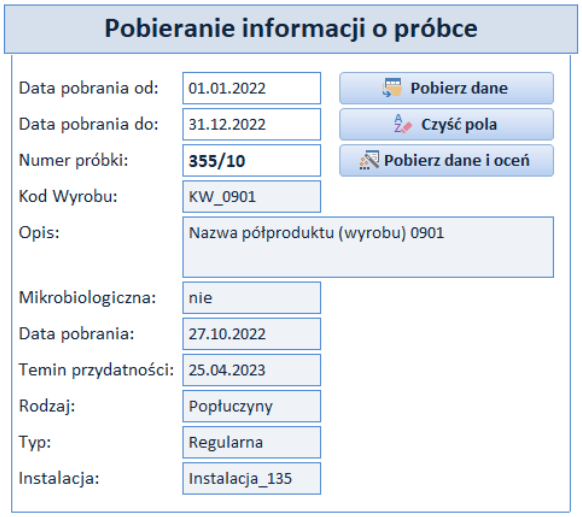
W celu poddania nowej partii wyrobu klasyfikacji, w formularzu wprowadza się jej identyfikator do pola „Numer próbki”. Zakład produkcyjny stosował numerowanie próbek w formacie: [liczba porządkowa] / [numer miesiąca]. Oznacza to, że próbka o numerze 355/10 była 355. próbką pobraną w październiku. Taki sposób identyfikacji nie zawiera informacji o roku, w którym nastąpiło przypisanie oznaczenia. Jednakże w bazie danych występuje dokładna data pobrania, dzięki czemu istniej możliwość ograniczenia zakresu wyszukiwania do analizowanego roku – pola ograniczania daty pobrania od/do (Rysunek 5.3).



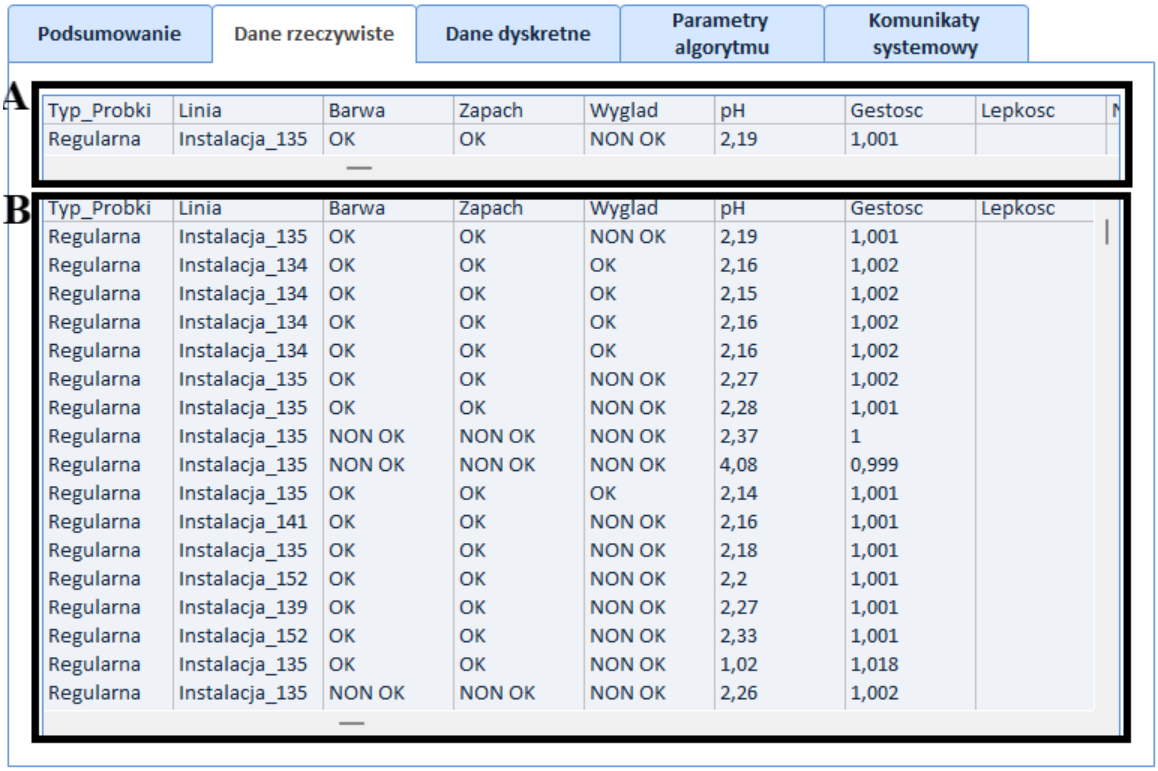
Rysunek 5.3. Opracowane narzędzie „Statystyczny Chemik” – pobieranie danych o próbce.

Na Rysunku 5.3 widoczne są trzy przyciski. Kolejno od góry służą one: (i) wyłącznie pobieraniu danych próbki i pozycji archiwalnych, (ii) usuwaniu wszelkich informacji z formularzy, (iii) pobraniu wszystkich danych, dyskretyzacji zmiennych rzeczywistych oraz dokonywaniu klasyfikacji. Pola zacieniowane nie mogą być edytowane. Po wskazaniu numeru identyfikacyjnego łącznie z zakresem dat, możliwe jest ściągnięcie informacji pierwotnych o próbce. W pierwszym kroku aplikacja uzyskuje parametry pozycji o wskazanym identyfikatorze. Następnie pobierane są informacje historyczne próbek odpowiadające temu samemu półproduktowi wraz z decyzją (etykietą). Jako wyróżnik do przeszukiwania zbioru danych służy kod wyrobu uzyskany w pierwszym w kroku. Obie czynności wykonują się po naciśnięciu jednego z przycisków pobierania danych.

Prezentację informacji dla próbki 355/10 (2022 rok) przedstawiono na Rysunku 5.4 (dane podstawowe) oraz na Rysunku 5.5 (dane rzeczywiste). Na drugiej ilustracji widnieje 8 parametrów z 17 dostępnych. W programie możliwe jest przesuwanie widoku, aby uzyskać podgląd wszystkich kolumn oraz wierszy.



Rysunek 5.4. Opracowane narzędzie „Statystyczny Chemik" – informacje podstawowe dla próbki o identyfikatorze 355/10 z 2022 roku.



A

Typ_Probki	Linia	Barwa	Zapach	Wyglad	pH	Gestosc	Lepkosc
Regularna	Instalacja_135	OK	OK	NON OK	2,19	1,001	

B

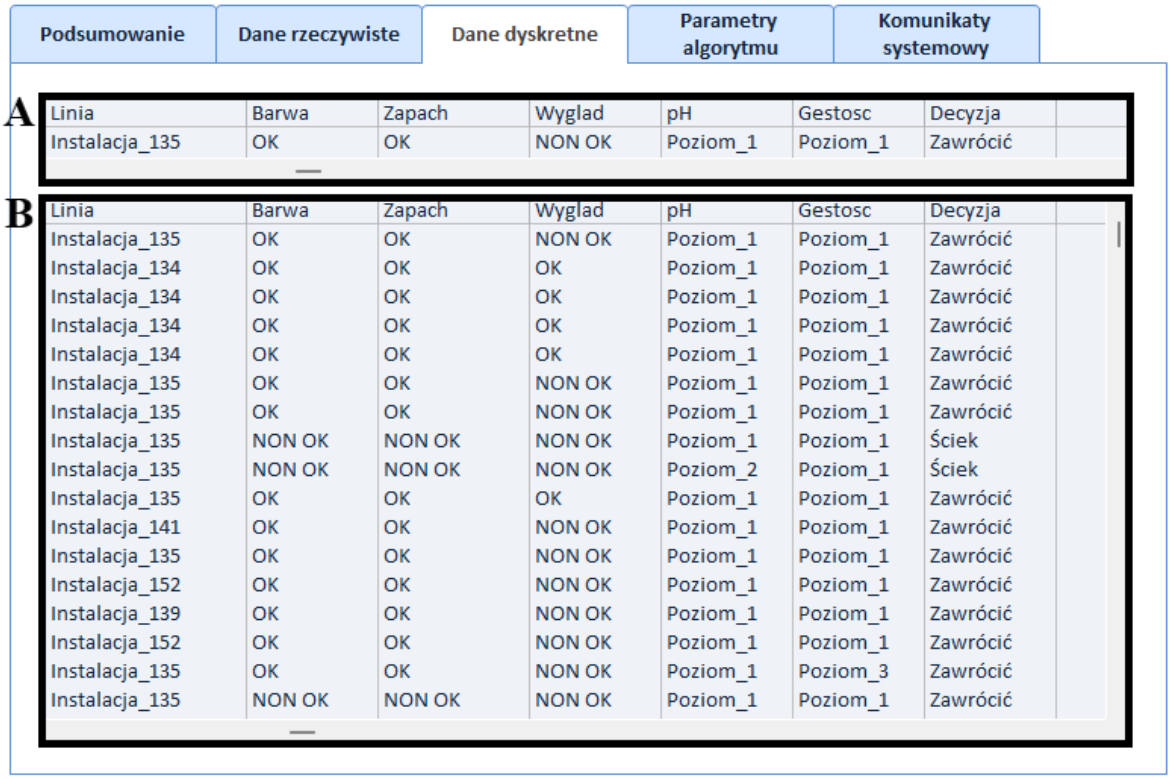
Typ_Probki	Linia	Barwa	Zapach	Wyglad	pH	Gestosc	Lepkosc
Regularna	Instalacja_135	OK	OK	NON OK	2,19	1,001	
Regularna	Instalacja_134	OK	OK	OK	2,16	1,002	
Regularna	Instalacja_134	OK	OK	OK	2,15	1,002	
Regularna	Instalacja_134	OK	OK	OK	2,16	1,002	
Regularna	Instalacja_134	OK	OK	OK	2,16	1,002	
Regularna	Instalacja_135	OK	OK	NON OK	2,27	1,002	
Regularna	Instalacja_135	OK	OK	NON OK	2,28	1,001	
Regularna	Instalacja_135	NON OK	NON OK	NON OK	2,37	1	
Regularna	Instalacja_135	NON OK	NON OK	NON OK	4,08	0,999	
Regularna	Instalacja_135	OK	OK	OK	2,14	1,001	
Regularna	Instalacja_141	OK	OK	NON OK	2,16	1,001	
Regularna	Instalacja_135	OK	OK	NON OK	2,18	1,001	
Regularna	Instalacja_152	OK	OK	NON OK	2,2	1,001	
Regularna	Instalacja_139	OK	OK	NON OK	2,27	1,001	
Regularna	Instalacja_152	OK	OK	NON OK	2,33	1,001	
Regularna	Instalacja_135	OK	OK	NON OK	1,02	1,018	
Regularna	Instalacja_135	NON OK	NON OK	NON OK	2,26	1,002	

Rysunek 5.5. Opracowane narzędzie „Statystyczny Chemik" – dane rzeczywiste dla próbki o identyfikatorze 355/10 z 2022 roku (ramka A) oraz dane historyczne odpowiadające temu samemu półproduktowi (ramka B).

5.3.2. Dyskretyzacja zmiennych rzeczywistych

Parametry rzeczywiste wskazane w Tabeli 5.1 (na stronie 66) poddawane są dyskretyzacji. Wbudowano w aplikację dwie metody transformacji liczb rzeczywistych do postaci dyskretnej: podział na równe przedziały oraz grupowanie oparte o odchylenie standardowe (techniki opisane zostały na stronie 36). Procesowi temu podlegają dane próbki klasyfikowanej oraz informacje o pozycjach historycznych. Wybór sposobu dyskretyzacji jest jednym z argumentów sterujących algorytmem klasyfikacji i podlegał on ewaluacji.

Niezależnie od wybranej metody algorytm w pierwszej kolejności weryfikuje, czy w danych historycznych argumenty posiadają wartości (np. pomiar lepkości nie jest wykonywany dla każdego półproduktu). Parametr będzie uwzględniany w obliczaniach, gdy w zbiorze wystąpi co najmniej jedna jego wartość. Natomiast, w ramach analizowanego argumentu próbkom z wartościami liczbowymi przyporządkowuje się poziomy dyskretne (ich liczba jest definiowana w ustawieniach algorytmu). W narzędziu zaprogramowano do wyboru dwie metody dyskretyzacji wartości rzeczywistych. Ponadto, zawsze dla tekstu przyznawany jest „Poziom_TXT", a dla braku jakiejkolwiek wartości „Poziom_PST". Prezentację danych po dyskretyzacji dla próbki 355/10 (2022 rok) przedstawiono na Rysunku 5.6.



Linia	Barwa	Zapach	Wyglad	pH	Gestosc	Decyzja
Instalacja_135	OK	OK	NON OK	Poziom_1	Poziom_1	Zawrócić

Linia	Barwa	Zapach	Wyglad	pH	Gestosc	Decyzja
Instalacja_135	OK	OK	NON OK	Poziom_1	Poziom_1	Zawrócić
Instalacja_134	OK	OK	OK	Poziom_1	Poziom_1	Zawrócić
Instalacja_134	OK	OK	OK	Poziom_1	Poziom_1	Zawrócić
Instalacja_134	OK	OK	OK	Poziom_1	Poziom_1	Zawrócić
Instalacja_134	OK	OK	OK	Poziom_1	Poziom_1	Zawrócić
Instalacja_135	OK	OK	NON OK	Poziom_1	Poziom_1	Zawrócić
Instalacja_135	OK	OK	NON OK	Poziom_1	Poziom_1	Zawrócić
Instalacja_135	NON OK	NON OK	NON OK	Poziom_1	Poziom_1	Ściek
Instalacja_135	NON OK	NON OK	NON OK	Poziom_2	Poziom_1	Ściek
Instalacja_135	OK	OK	OK	Poziom_1	Poziom_1	Zawrócić
Instalacja_141	OK	OK	NON OK	Poziom_1	Poziom_1	Zawrócić
Instalacja_135	OK	OK	NON OK	Poziom_1	Poziom_1	Zawrócić
Instalacja_152	OK	OK	NON OK	Poziom_1	Poziom_1	Zawrócić
Instalacja_139	OK	OK	NON OK	Poziom_1	Poziom_1	Zawrócić
Instalacja_152	OK	OK	NON OK	Poziom_1	Poziom_1	Zawrócić
Instalacja_135	OK	OK	NON OK	Poziom_1	Poziom_3	Zawrócić
Instalacja_135	NON OK	NON OK	NON OK	Poziom_1	Poziom_1	Zawrócić

Rysunek 5.6. Opracowane narzędzie „Statystyczny Chemik" – dane dyskretne dla próbki o identyfikatorze 355/10 z 2022 roku (ramka A) oraz dane historyczne odpowiadające temu samemu półproduktowi (ramka B).

Dyskretyzacja na równe przedziały

Metoda dyskretyzacji na równe przedziały została zaimplementowana zgodnie ze wzorem (8) – patrz strona 36. W zbiorze danych historycznych dla każdego parametru wskazywany jest zakres liczbowy od jego wartości minimalnej do maksymalnej. Następnie jest on dzielony na równe przedziały. W ten sposób uzyskiwane są zmienne dyskretne, czyli poziomy, które są oznaczane kolejnymi numerami porządkowymi.

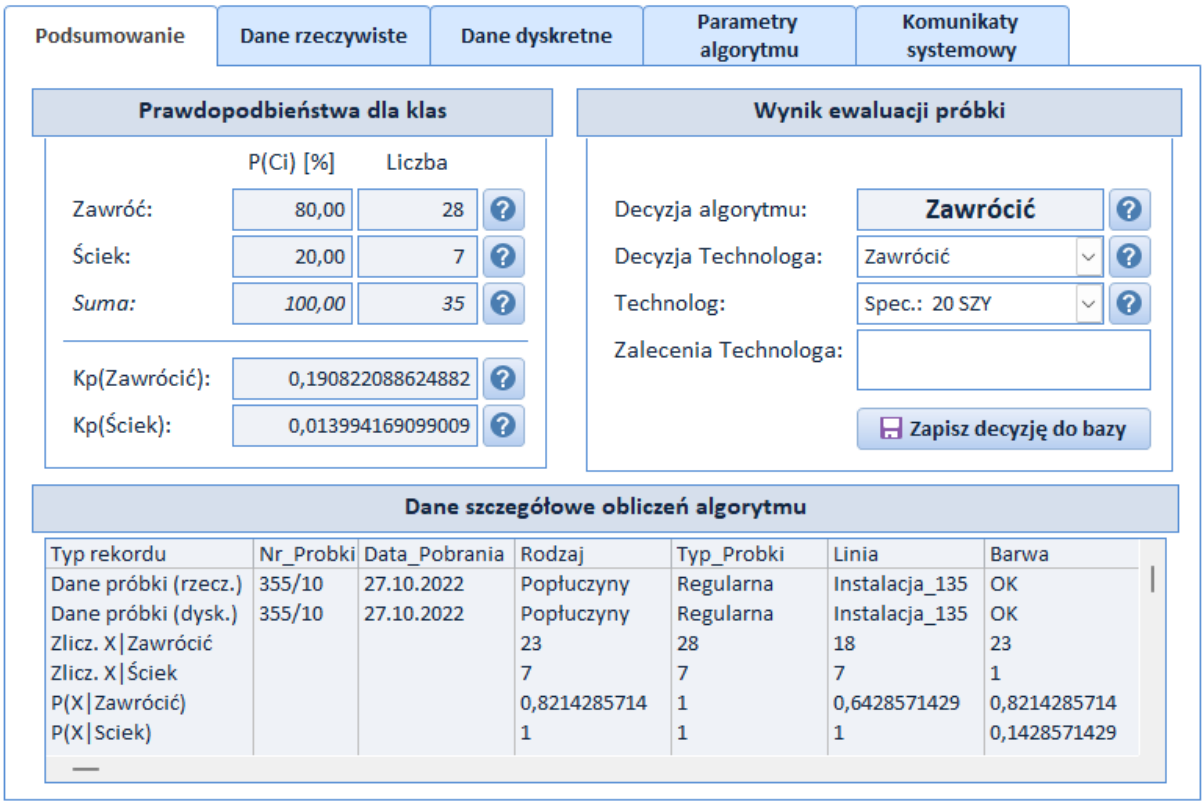
Dyskretyzacja oparta o odchylenie standardowe

W aplikacji zaimplementowano również dyskretyzację, która wykorzystuje wartość odchylenia standardowego każdego argumentu. W tej metodzie przyjęto, że granice przedziału będą przyjmować wartość średnią pomniejszoną (dolna granica) lub powiększoną (górna granica) o trzy odchylenia standardowe ($\bar{x} \pm 3s$). Tak więc, przedział o szerokości sześciu odchyleń standardowych jest dzielony na zdefiniowaną przez użytkownika liczbę zakresów. Jeżeli liczba ta zostanie określona na 3, to dyskretyzacja będzie zgodna z graficzną interpretacją przedstawioną na Rysunku 4.7 – patrz strona 36. Istotne jest również uwzględnienie wartości, które wykraczają poza granice przedziału, ponieważ zgodnie z teorią rozkładu normalnego w tym zakresie nie występują wszystkie możliwe obserwacje. W takim przypadku zostanie przypisany „Poziom_URN”, gdy wartość będzie mniejsza od dolnej granicy, lub „Poziom_ARN” w sytuacji przekroczenia górnej wartości zakresu.

5.3.3. Przypisywanie próbki do klasy

Celem narzędzia jest przypisanie badanej próbce klasy – należy wprowadzić identyfikator próbki oraz użyć przycisku „Pobierz dane i oceń” (Rysunek 5.3). W ten sposób uruchamiana jest procedura: pobierania danych pierwotnych, transformacji na poziomy dyskretne oraz obliczenia prawdopodobieństw zgodnie z metodą naiwnej klasyfikacji Bayesa.

Wynik pracy algorytmu prezentowany jest na karcie podsumowania (Rysunek 5.7). W obszarze „Prawdopodobieństwo dla klas” wyświetlane są parametry odpowiadającej każdej z etykiet, jest to: liczba wystąpień w zbiorze treningowym oraz odpowiadające jej prawdopodobieństwo. Współczynniki *Kp (Zawrócić)* oraz *Kp (Ściek)* stanowią licznik ze wzoru (6) – patrz strona 35. Klasa, dla której zostanie osiągnięta większa wartość współczynnika zostaje przypisana jako decyzja algorytmu – zgodnie z teorią opisaną w Rozdziale 4.3.6. Dolna ramka zawiera listę szczegółowych informacji oraz wyników jakie były wykorzystywane przez algorytm w celu dokonania klasyfikacji. W obszarze „Wynik ewaluacji próbki” algorytm wyświetla informację o przypisanej klasie, jeżeli przyznanie etykiety nie było możliwe, to wyświetlana jest czerwona ikona. Zgodnie z wymaganiami przedsiębiorcy (Tabela 5.2) trzy jasne pola służą personelowi do zapisania ostatecznej decyzji.

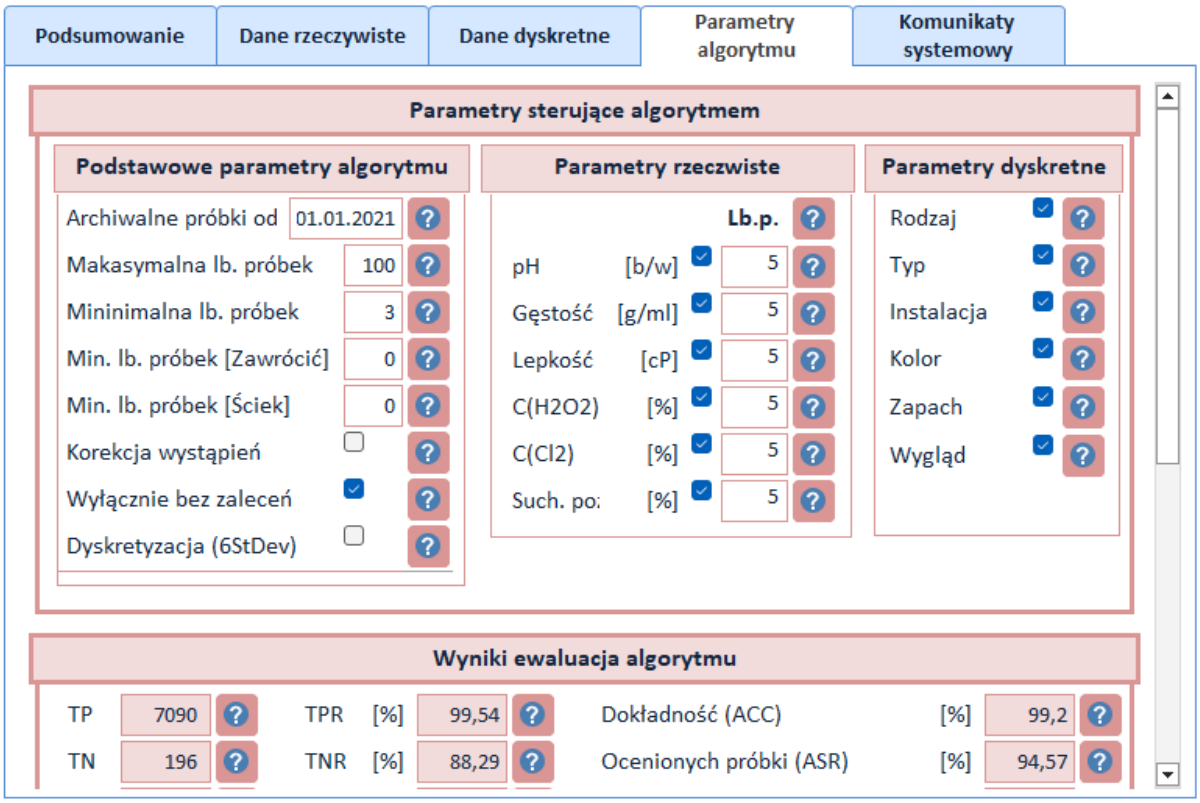


Rysunek 5.7. Opracowane narzędzie „Statystyczny Chemik” – podsumowanie klasyfikacji dla próbki o identyfikatorze 355/10 z 2022 roku.

5.3.4. Możliwości konfiguracyjne algorytmu

Na Rysunku 5.8 przedstawiono kartę służącą parametryzacji oraz ewaluacji algorytmu klasyfikacji. W pierwszej ramce o nagłówku „Parametry sterujące algorytmem" zebrane są trzy grupy zmiennych, które wpływają na działanie uczenia maszynowego. Konfiguracja aplikacji następuje przez zmianę wartości prezentowanych ustawień.

W pierwszej sekcji „Podstawowe parametry" zestawiono zmienne, które nie charakteryzują próbek produkcyjnych, opisano je szerzej w Tabeli 5.3. W środkowej ramce prezentowane są „Parametry rzeczywiste", obejmujące wyniki analiz fizykochemicznych. Przyjmują one wartości liczb rzeczywistych. Konfigurowane są w dwoisty sposób: poprzez oznaczenie ich jak nieaktywne (nieuwzględnienie w obliczaniu prawdopodobieństw) oraz przez definicję liczby przedziałów stosowanych w procesie dyskretyzacji. W Tabeli 5.3 przedstawiono definicję skróconych nazw stosowanych w aplikacji oraz na prezentowanym Rysunek 5.8. Ostatnia grupa zawiera parametry dyskretne. Ich nazewnictwo jest zgodne tym przedstawionym w Tabeli 5.1. Parametryzacja algorytmu z użyciem tej grupy polega na włączeniu bądź nie rozpatrywanego atrybutu do obliczeń klasyfikacji.



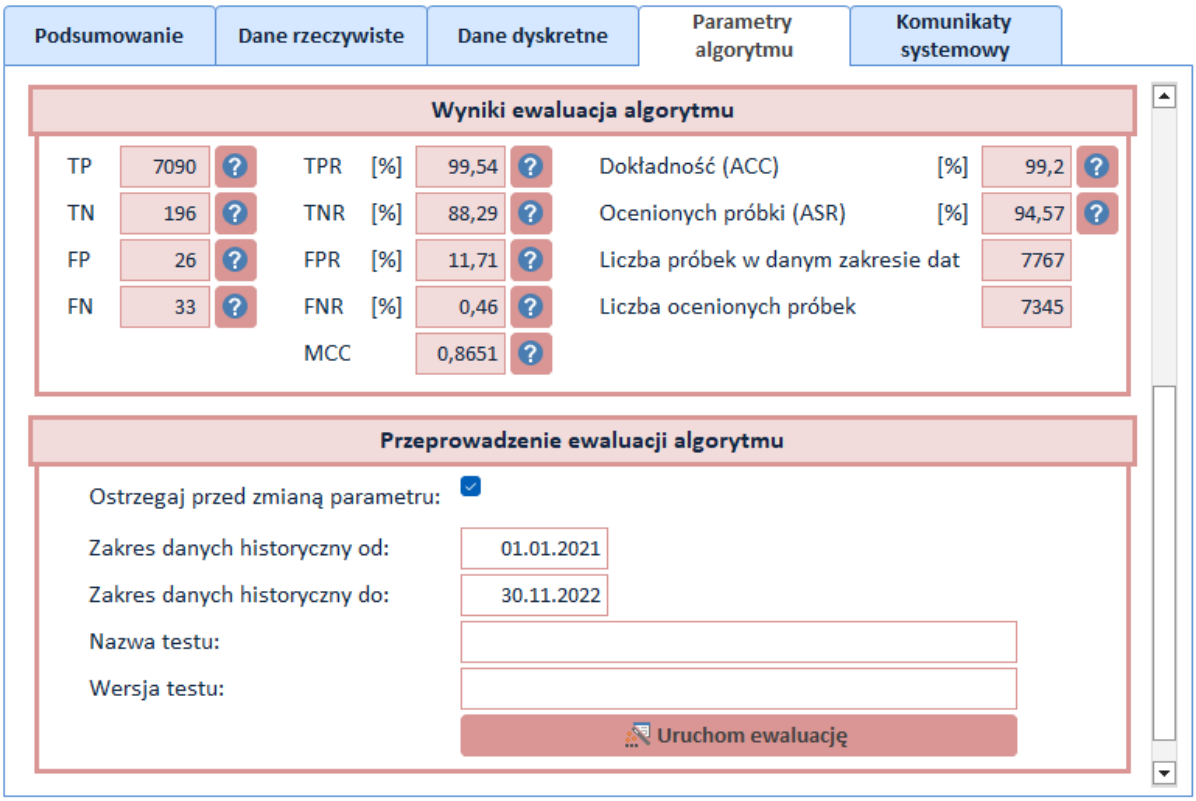
Rysunek 5.8. Opracowane narzędzie „Statystyczny Chemik" – parametry sterujące algorytmem.

Tabela 5.3. Opracowane narzędzie „Statystyczny Chemik” – opis skróconych nazw parametrów sterujących algorytmem.

	Nazwa parametru używana w aplikacji	Opis parametru
Podstawowe parametry algorytmu	Archiwalne próbki od	Ograniczenie liczby próbek historycznych (zbioru treningowego). Jest to data od której mogą być stosowane zapisy archiwalne.
	Maksymalna lb. próbek	Maksymalna liczba próbek pobierana z bazy danych w kolejności od najmłodszej do najstarszej.
	Minimalna lb. próbek	Minimalna liczba próbek treningowych, wymagana w celu dokonania klasyfikacji. W przypadku, jeżeli nie jest ona osiągnięta, algorytm nie dokona klasyfikacji.
	Min. lb. próbek [Zawrócić]	Wymagana minimalna liczba próbek archiwalnych, odpowiednio oznaczonych jako Ściek oraz Zawrócić w celu dokonania klasyfikacji.
	Min. lb. próbek [Ściek]	
	Korekcja wystąpień	Jeżeli parametr jest zaznaczony to stosowana jest korekcja zerowego iloczynu prawdopodobieństw. Stosowane jest wygładzanie Laplace'a – zwiększanie licznika i mianownika o jeden w obliczaniach prawdopodobieństwa dla każdego atrybutu [154].
	Wyłącznie bez zaleceń	Jeżeli parametr jest zaznaczony to w celu obliczeń klasyfikacji pobierane są z bazy danych (zbioru treningowego) próbki wyłącznie bez zaleceń technologa.
	Dyskretyzacja (6StDev)	Jeżeli parametr jest zaznaczony to dyskretyzacja odbywa się w oparciu o odchylenie standardowe, w przeciwnym razie stosowany jest podział na równe przedziały.
Parametry rzeczywiste	pH	Bezwymiarowa wartość z przedziału od 1 do 14.
	Gęstość	Wynik pomiaru gęstości wyrażony w gramach na mililitr.
	Lepkość	Wartość lepkości zbadana z wykorzystaniem lepkościomierzy Brookfielda wyrażona w centypuazach.
	C(H2O2)	Wynik oznaczenia stężenia procentowego nadtlenku wodoru.
	C(CL2)	Oznaczenie wolnego chloru jako stężenie procentowe.
	Such. po.	Sucha pozostałość wyrażona w procentach masowych.

5.3.5. Ewaluacja parametrów konfiguracyjnych algorytmu

W narzędziu zostały zaimplementowane metody ewaluacji algorytmu. Na Rysunku 5.9 pokazano elementy oceny klasyfikacji. Umieszczono je w dolnej części karty parametrów algorytmu. W pierwszej (górnej) ramce wyświetlane są wskaźniki, które wybrano w celu bieżącego monitorowania działania narzędzia. Parametry te zostały opisane w Rozdziale 4.4 (strona 40). Ponadto prezentowana jest informacja, jaka część próbek spośród dostępnych w bazie danych treningowych została sklasyfikowana przez narzędzie (przyznanie klasy „Zawrócić" lub „Ściek"). Informacja ta wyświetlana jest w postaci procentowej (ASR) oraz wartości liczb bezwzględnych.



Rysunek 5.9. Opracowane narzędzie „Statystyczny Chemik" – elementy ewaluacji algorytmu.

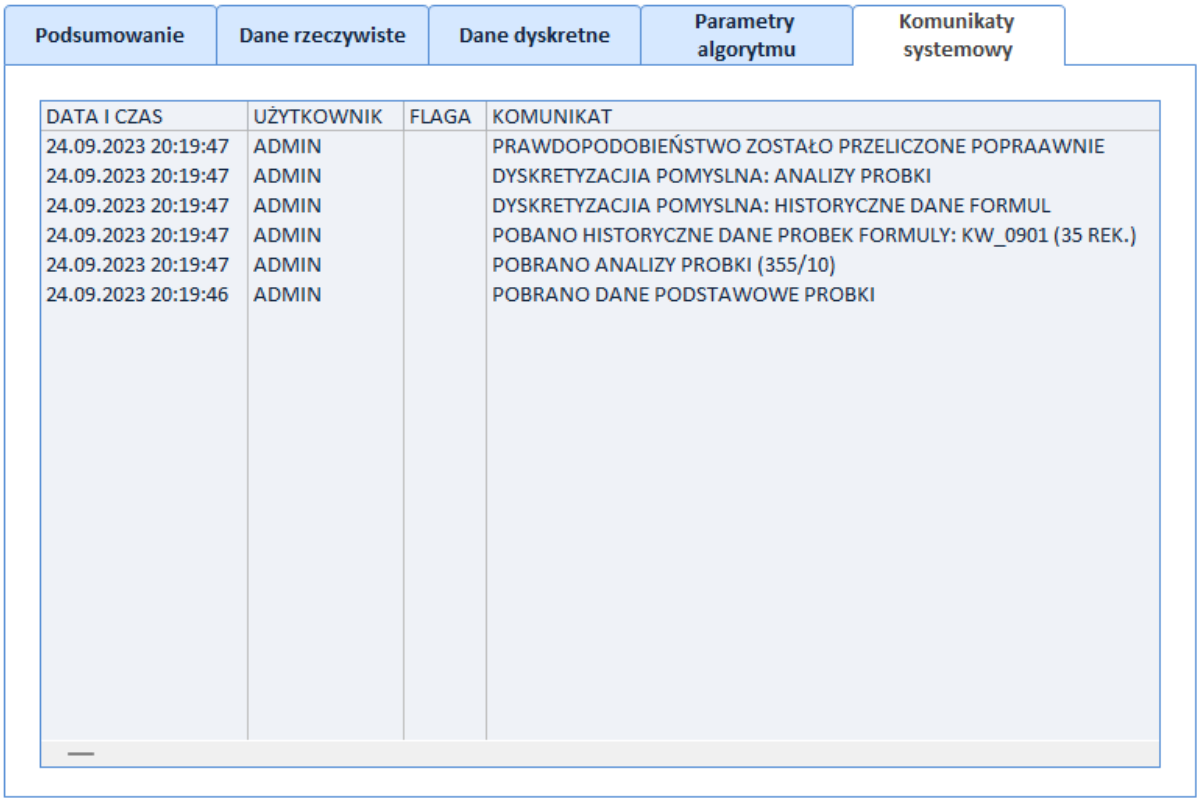
Na Rysunku 5.9 w obszarze o nagłówku „Przeprowadzanie ewaluacji algorytmu" (dolna ramka) znajduje się część funkcjonalność, która służy użytkownikowi do weryfikacji nastaw narzędzia. Każda zmianie parametru ustawień (opisanych na stronie 75) powoduje resetowanie wyników ewaluacji. W celu uzyskania metryk skuteczności algorytmu należ wybrać zakres dat zbioru treningowego (daty pobrania próbek produkcyjnych) oraz uruchomić ewaluację. Możliwe jest nadanie uruchamianej ewaluacji nazwy oraz tytułu wersji.

W celu minimalizacji ryzyka oraz zgodnie z wymaganiami przedsiębiorcy (Tabela 4.1) wdrożony został rejestr zmian nastaw algorytmu. Jeżeli nastąpi aktualizacja któregokolwiek parametru sterującego klasyfikacją, to narzędzie wykona dwie czynności: (i) zapisze w rejestrze wartości nastaw przed modyfikacją, wyniki ewaluacji oraz datę zmiany, (ii) wygeneruje plik tekstowy, który będzie zawierał dane analogiczne do tych w rejestrze, a ponadto informacje o każdej ocenionej próbce produkcyjnej. Aby narzędzie wskazywało aktualne dane, to wskaźniki oceny ewaluacji aktualizowane są po każdym zapisaniu próbki przez personel.

5.3.6. Komunikaty oraz obsługa błędów

Komunikaty systemowe

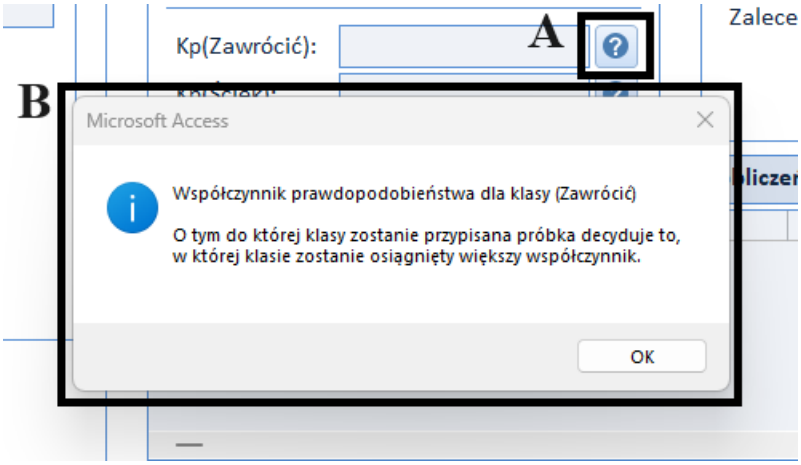
W celu nadzorowania pracy narzędzia zaprojektowana została karta komunikatów systemowych (Rysunek 5.10). Po każdym zakończonym etapie pracy algorytmu do listy dodawana jest pozycja o powodzeniu operacji bądź o jej błędzie. W przypadku zaistnienia nieprawidłowości, historia wpisów w tym oknie umożliwia identyfikacje przyczyny.



Rysunek 5.10. Opracowane narzędzie „Statystyczny Chemik” – komunikaty systemowe dla próbki o identyfikatorze 355/10 z 2022 roku.

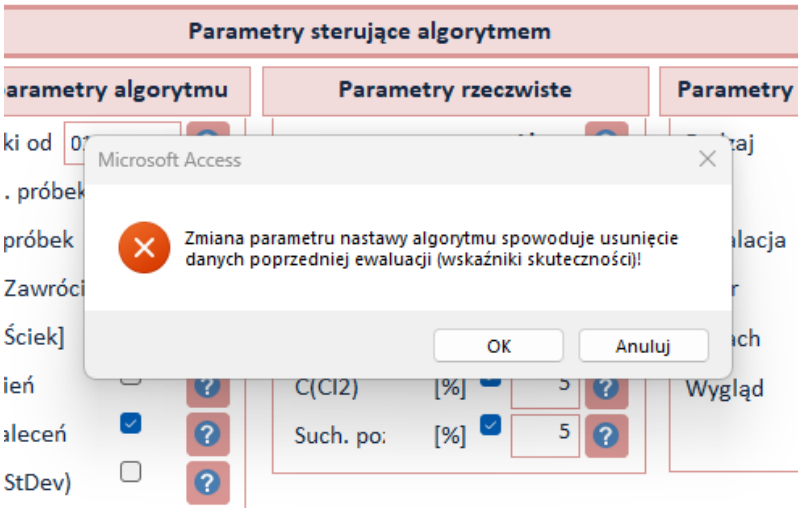
Okienka informacyjne

W interfejsie graficznym zostały umieszczone przyciski oznaczone znakiem zapytania. Służą one wywoływaniu informacji o polach oraz wartościach w nich przechowywanych. Na Rysunku 5.11 przedstawiony został przykładowy komunikat (ramka B) wywoływany przez przycisk (ramka A) przy współczynniku prawdopodobieństwa.



Rysunek 5.11. Opracowane narzędzie „Statystyczny Chemik" – ilustracja informacji kontekstowa o polach interfejsu: ramka A – przycisk wywoływania, ramka B – wywoływana informacja.

W ustawieniach parametrów sterujących algorytmem zostało wdrożone zabezpieczenie w postaci ostrzeżenia, które ma na celu zapobiec przypadkowym modyfikacjom. Nieumyślna zmiana mogłaby skutkować utraceniem nastaw oraz wartości wskaźników oceniających. Użytkownikowi udostępniono opcję anulowania zmian (Rysunek 5.12).



Rysunek 5.12. Opracowane narzędzie „Statystyczny Chemik" – ostrzeżenie wywoływane zmianą parametrów algorytmu.

5.4. Model testowania klasyfikacji

Ostatnim etapem realizacji badań było zmierzenie wpływu poszczególnych parametrów na skuteczność klasyfikacji. W tym celu przeprowadzono szereg testów, które w sposób iteracyjny modyfikowały wartość każdej nastawy względem punktu odniesienia. Uzyskano w ten sposób charakterystyki pokazujące wpływ badanych parametrów na kluczowe wskaźniki oceny algorytmu. W Tabeli 5.4 zestawiono wartości ustawień, które zostały wskazane jako punkt odniesienia wraz z uzasadnieniem. Te wartości również zaprezentowano na Rysunku 5.8, natomiast wyniki ich ewaluacji pokazano na Rysunku 5.9.

Tabela 5.4 zawiera także plan, według którego poddano optymalizacji nastawy. W ogólności, parametry rzeczywiste testowano poprzez przyrostową zmianę ich wartości. Natomiast atrybuty dyskretne połączono w grupy, które były weryfikowane w procedurze kombinatorycznej. Nie było możliwe przetestowanie wszystkich parametrów dyskretnych (15 pozycji) w jednej grupie, ponieważ liczba możliwości ich konfiguracji wynosi 32 768 (2^{15}). Po każdej modyfikacji uruchamiano analizę zbioru treningowego, której efektem była aktualizacja wskaźników oceny klasyfikacji. Następnie na podstawie wartości wyników ewaluacji z każdej grupy testów, wszystkie atrybuty zostały wysterowane jednocześnie, w celu zweryfikowania czy wystąpi efekt synergii.

Ostatnią fazą testów było sprawdzenie czy narzędzie może zostać wykorzystane w celu ograniczenia kosztów kontroli jakości. Taka sytuacja będzie miała miejsce, jeżeli kryteria sukcesu (opisane na stronie 68) zostaną osiągnięte z pominięciem badań kosztochłonnych, czyli takich do wykonania których zużywane są odczynniki oraz materiały jednorazowe. W rozpatrywanym przypadku będą to: pomiar suchej pozostałości, oznaczanie stężenia wolnego chloru oraz nadtlenku wodoru.

Tabela 5.4. Wartości parametrów algorytmu w punkcie odniesienia oraz plan testów optymalizacyjnych.

	Nazwa parametru używana w aplikacji	Wartość parametru	Uzasadnienie przyjętej wartości parametru w punkcie odniesienia	Plan testowania parametru
Podstawowe parametry algorytmu	Archiwalne próbki od	01.01.2021	Średni okres pomiędzy modyfikacją półproduktu to około 2 lata. Ponieważ miała miejsce aktualizacja receptur, wyroby sprzed 2021 roku nie powinny być uwzględniane.	Nie podlega zmianom optymalizacyjnym.
	Maksymalna lb. próbek	100	Zastosowano w celu redukcji obciążania komputera. Ponad 75% półproduktów posiada mniej niż 100 próbek, zaś średnia arytmetyczna wynosi 78.	Przyrostowa zmiana wartości parametru. Od 3 do 25 zmiana wartości o 1, powyżej 50 przyrost o 25. Sumarycznie 45 testów.
	Minimalna lb. próbek	3	W przypadku minimalnej liczby 3 próbek będzie przeważała jedna klasa.	Przyrostowa zmiana wartości parametru o 1 począwszy od 0. Koniec testów wyznacza wartość wskaźnika liczby sklasyfikowanych próbek (ASR) – poniżej 80%.
	Min. lb. próbek [Zawrócić]	0	Wartość minimalna nie została zdefiniowana w celu niezaburzania proporcji pomiędzy klasami w zbiorze treningowym.	
	Min. lb. próbek [Ściek]	0		
	Korekcja wystąpień	Wyłączona	W punkcie odniesienia uwzględniano jedynie parametry, które są powiązane bezpośrednio ze zbiorem danych.	Grupa parametrów dyskretnych testowana w sposób kombinatoryczny. Istnieje 8 możliwości: 3 parametry, które mogą przyjąć 2 stany (włączony lub wyłączony).
	Wyłącznie bez zaleceń	Włączona	Aby nie zaburzać klasyfikacji binarnej, próbki z zaleceniami (zawrócone warunkowo) nie są uwzględniane.	
	Dyskretyzacja (6StDev)	Wyłączona	W celu redukcji obciążania komputera domyślnie stosowana jest dyskretyzacja na równe przedziały.	

	Nazwa parametru używana w aplikacji	Wartość parametru	Uzasadnienie przyjętej wartości parametru w punkcie odniesienia	Plan testowania parametru
Parametry rzeczywiste	pH	Włączone, 5 poziomów dyskretnych.	W celu wykorzystania wszystkich analiz fizykochemicznych wyznaczających jakość wyrobu, uwzględniono w punkcie odniesienia wszystkie parametry. Zastosowano 5 poziomów dyskretny, by ograniczyć obciążenie komputera.	Każdy parametr testowany niezależnie. Zmiana aktywności testowanego atrybutu, a następnie przyrostowa modyfikacja liczby poziomów dyskretnych, zgodnie z planem: - od 1 do 50 zmiana wartości o 1, - od 55 do 150 zmiana wartości o 5, - od 160 do 200 zmiana wartości o 10, - od 300 do 1500 zmiana wartości o 100. 90 iteracji każdego parametru, sumarycznie 540 testów.
	Gęstość			
	Lepkość			
	C(H2O2)			
	C(CL2)			
	Such. po.			
Parametry dyskretne	Rodzaj	Włączone	Wszystkie atrybuty są cechą danej partii ocenianego wyrobu, mogące mieć wpływ na wynik klasyfikacji.	Grupa parametrów dyskretnych testowana w sposób kombinatoryczny. Istnieje 8 możliwości: 3 parametry, które mogą przyjąć 2 stany (włączony lub wyłączony).
	Typ			
	Instalacja			
	Kolor	Włączone	W celu wykorzystania wszystkich analiz fizykochemicznych wyznaczających jakość wyrobu, uwzględniono w punkcie odniesienia wszystkie parametry.	Grupa parametrów dyskretnych testowana w sposób kombinatoryczny. Istnieje 8 możliwości: 3 parametry, które mogą przyjąć 2 stany (włączony lub wyłączony).
	Zapach			
	Wygląd			

6. Wyniki badań

Celem rozprawy było zbadanie możliwości wykorzystania narzędzi statystycznych w analizach jakościowych (obszary produkcyjne) w przemyśle chemicznym. W tym celu podjęta została współpraca z przedsiębiorstwem, które wytwarzało około 1 mln opakowań artykułów chemii gospodarczej w trakcie doby. Realizowany szarżowy proces wytwórczy (opisany na stronie 58) zawierał w swoich ramach działania kontroli jakości. Pracę badawczą podzielono na trzy etapy:

1) Opracowanie koncepcji algorytmu oraz aplikacji, która mogłaby być zastosowana w obszarze produkcyjnym – Rozdział 5.2.
2) Zbudowanie aplikacji w sposób pozwalający realizować przedmiot badań oraz zgodny z wymaganiami przedsiębiorstwa – Rozdział 5.3.
3) Badanie parametrów algorytmu oraz weryfikacja czy osiągnięte zostały kryteria sukcesu (opisane na stronie 68) – wyniki tego etapu przedstawiono w niniejszym rozdziale.

Realizacja pierwszych dwóch etapów pozwoliła na badania narzędzia statystycznego, wykorzystującego metody z zakresu uczenia maszynowego. Badano zmiany:

- nastaw modelu klasyfikatora niezwiązanych z właściwościami próbki,
- modyfikacje parametrów charakteryzujących próbkę,
- wszystkich parametrów, w ramach najlepszych odpowiedzi z poprzednich grup testowych,
- wykorzystania kosztochłonnych analiz fizykochemicznych.

Każdy parametr był oceniany przy użyciu minimum 7767 próbek pochodzących ze zbioru treningowego. Uzyskane wyniki przedstawiono w niniejszej części pracy.

6.1. Ogólny punkt odniesienia

Pojedynczy test wartości parametrów, określanych jako punkt odniesienia (referencyjny) oznaczono identyfikatorem T221, wyniki jego ewaluacji zaprezentowano w Tabeli 6.1.

Tabela 6.1. Wyniki ewaluacji – ogólny punkt odniesienia.

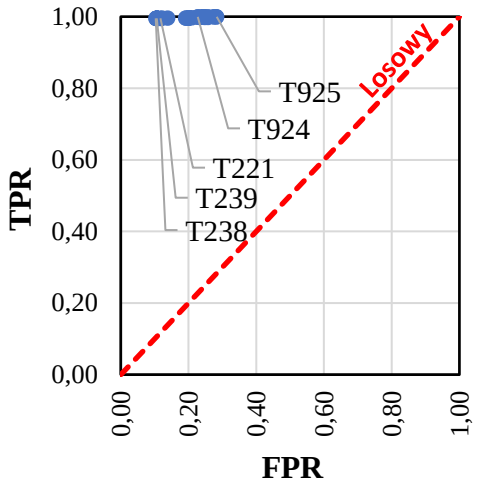
Numer testu	MCC	ACC	ASR	TPR	FPR	TNR	TP	FP
T221	0,8651	99,20%	94,57%	99,54%	11,71%	88,29%	7090	26

Jako punkt odniesienia dobrano arbitralnie wartości nastaw (Tabela 5.4). Charakteryzują się one wysokimi wskaźnikami jakości klasyfikacji. Uzyskana dokładność (ACC) wyniosła powyżej 99%. Zauważalny jest jednak wskaźnik próbek fałszywie pozytywnych (FPR), który osiągnął poziom powyżej 11%. Liczba próbek zaklasyfikowanych błędnie do zawrócenia (FP = 26) wskazuje na dużą asymetrię pomiędzy licznością klas. Współczynnik korelacji (MCC), który uwzględnia niezbalansowanie zbioru treningowego wyniósł 0,8651 (wartość 1 oznacza idealną korelację).

6.2. Modyfikacja parametrów niezwiązanych z właściwościami próbki

6.2.1. Maksymalna ogólna liczba próbek uwzględniana do obliczeń

W pierwszej kolejności zmieniana była maksymalna ogólna liczba próbek, uwzględnianych w obliczeniach prawdopodobieństwa. W Tabeli 6.2 zestawione są te testy, dla których rozpatrywane wskaźniki oceny osiągnęły swoje maksimum, łącznie z wartością parametru, która była poddana ewaluacji. Na Rysunku 6.1 naniesiono wszystkie punkty testowe, zaś etykietą oznaczono jedynie warianty przedstawione w Tabeli 6.2. Wartość TPR pozostaje na stale wysokim poziome. Obserwowana jest znacząca poprawa FPR wraz ze wzrostem limitu – czyli maleje liczba próbek sklasyfikowanych fałszywie pozytywnie. Obszar powyżej czerwonej linii oznacza pozytywną korelację wyników klasyfikacji z etykietami zawartymi w zbiorze treningowym. Klasyfikator idealny (ACC = 100%) przyjmuje wartości TPR = 1 oraz FPR = 0. Punkty znajdujące się na linii „Losowy” (TPR = FPR) charakteryzuje czysto losowe prawdopodobieństwo przypisania klasy. Obszar pod linią oznacza negatywną korelację wyników algorytmu względem wartości rzeczywistych ze zbioru danych uczących.

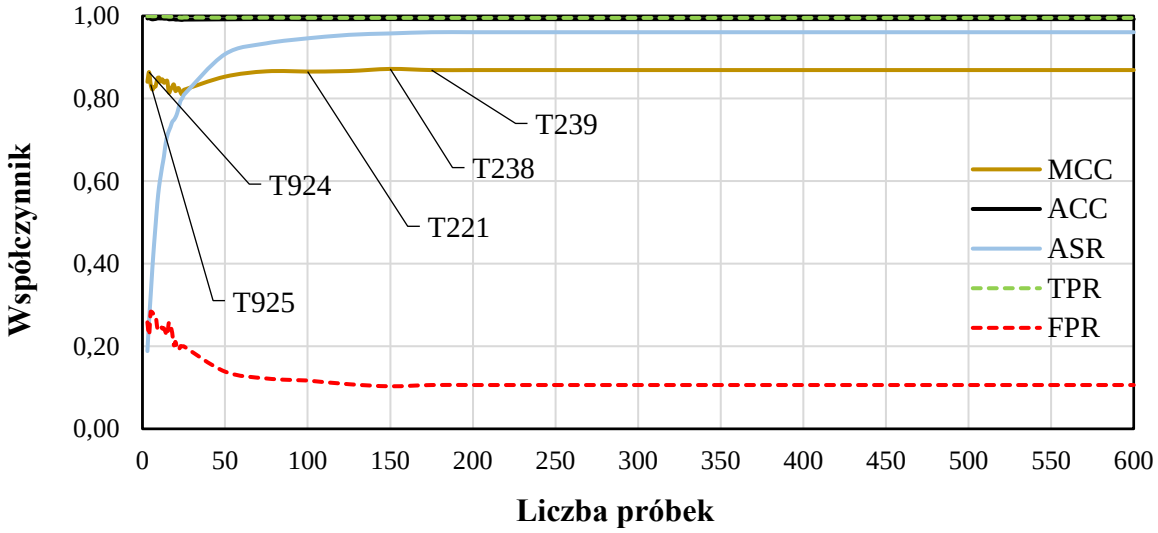


Rysunek 6.1 Układ współrzędnych TPR-FPR – wyniki ewaluacji zmiany parametru maksymalnej ogólnej liczby próbek.

Tabela 6.2. Zestawienie testów, w których współczynnik osiągnął maksimum – ogólna liczba próbek.

Numer testu	Zmaksymalizowany współczynnik	Wartość nastawy
T924	ACC	4
T238	MCC	150
T239	ASR	175
T925	TPR	5
T238	TNR = (1 - FPR)	5
T221		100

Wartość dokładności osiągnęła w każdym przeprowadzonym teście wynik powyżej 99% (ACC_{MIN} = 99,03%; ACC_{MAX} = 99,46%). Wskaźnik TPR również we wszystkich wariantach osiągnął wartość powyżej 99%. Na Rysunku 6.2 krzywe TPR i ACC nakładają się na siebie. Istotna jest zmiana liczby próbek, którym przyznano jakąkolwiek klasę. Wraz z inkrementacją wartości parametru obserwowany jest znaczący wzrost wartości ASR. W zbiorze treningowym 80% próbek (6214 sztuk) otrzymało klasę dopiero, gdy limit osiągnął wartość 24. Dla maksymalnej ogólnej liczby próbek wyższej niż 100 wszystkie wskaźniki się stabilizują.

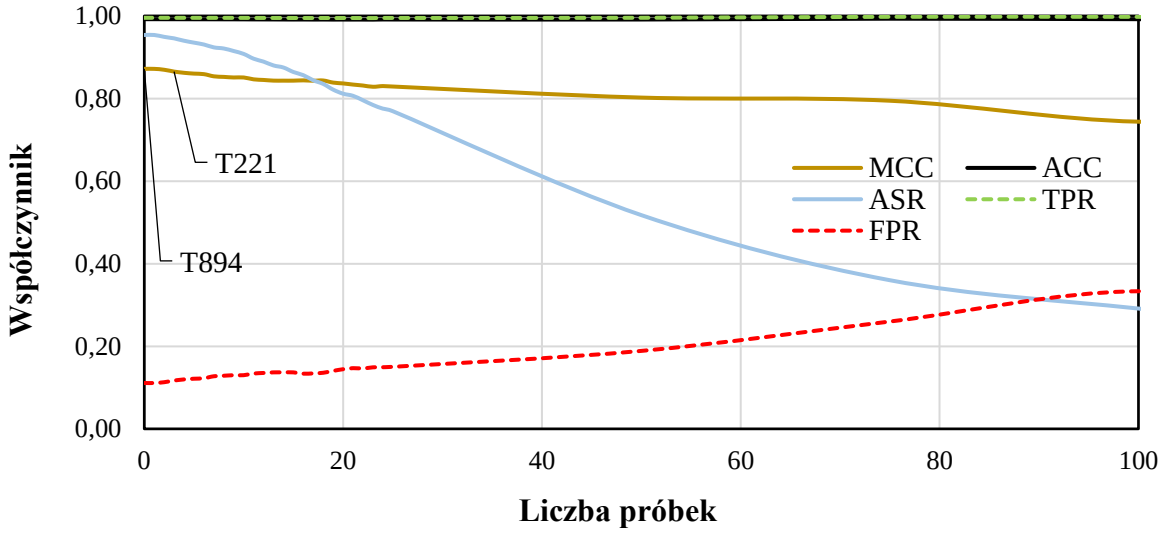


Rysunek 6.2. Profile wskaźników oceny klasyfikacji zarejestrowane podczas zmiany parametru maksymalnej ogólnej liczby próbek.

6.2.2. Wymagana minimalna ogólna liczba próbek

Badany parametr stanowi wymaganą minimalną liczbę próbek ze zbioru treningowego potrzebną do klasyfikacji. Jeżeli wartość ta nie będzie dostępna, narzędzie nie dokona klasyfikacji oraz wyświetli odpowiedni komunikat. Parametr był modyfikowany inkrementacyjnie do osiągniecia wartości 100. Już przy wartości 22 przekroczona została granica (ASR < 80%), która pozwalała przerwać optymalizację (zgodnie z planem testów).

Wraz ze wzrostem wartości parametru badanego, obserwowany jest spadek jakości klasyfikacji (Rysunek 6.3). Gwałtownie spada możliwość narzędzia do przyznania oceny próbkom w zbiorze treningowym (ASR). Z początkowej wartości 99,38% (0 poziomów) do 29,14% (100 poziomów). Zauważalna jest stabilna liczba próbek prawdzie pozytywnych (TPR). Jednocześnie zaobserwowano ciągły wzrost wskaźnika próbek fałszywie pozytywnych (FPR). Następuje pogorszenie klasyfikacji w ramach klasy negatywnej. Współczynnik korelacji (MCC) również systematycznie pogarszała się wraz z kolejnymi inkrementacjami parametru. Krzywe TPR i ACC nakładają się na siebie. Najlepsze oceny uzyskał test T894, w którym wartość argumentu wynosiła 0 (ACC = 99,20%; ASR = 95,38%; TPR = 99,54%; FPR = 11,11%; MCC = 0,8718). Różnica pomiędzy punktem odniesienia (T221) wynika z mniejszej liczby ocenionych próbek (ASR_{T221} = 94,57%).

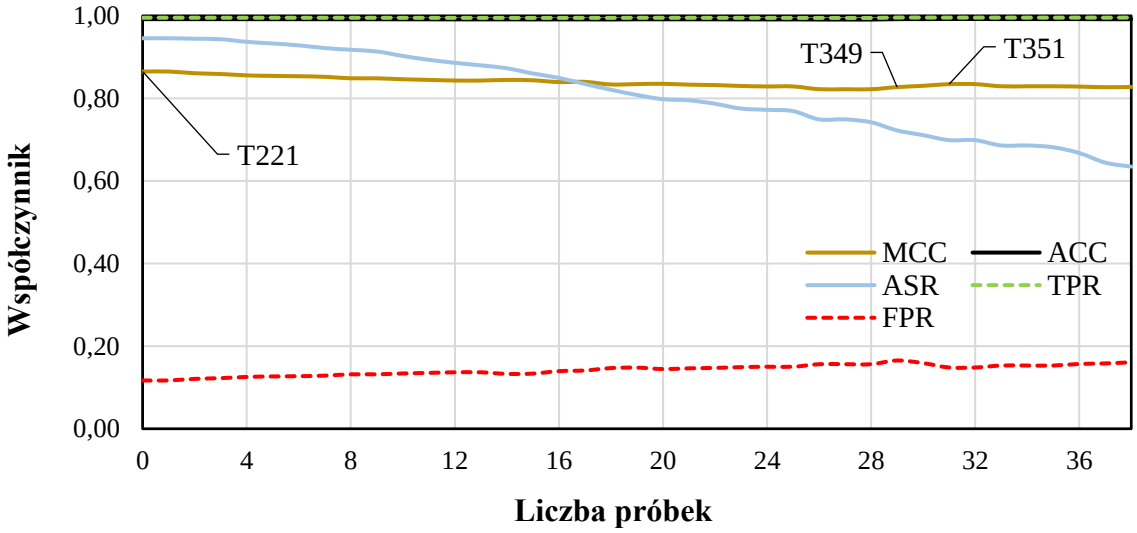


Rysunek 6.3. Profile wskaźników oceny klasyfikacji zarejestrowane podczas zmiany parametru minimalnej ogólnej liczby próbek.

6.2.3. Wymagana minimalna liczba próbek z klasy

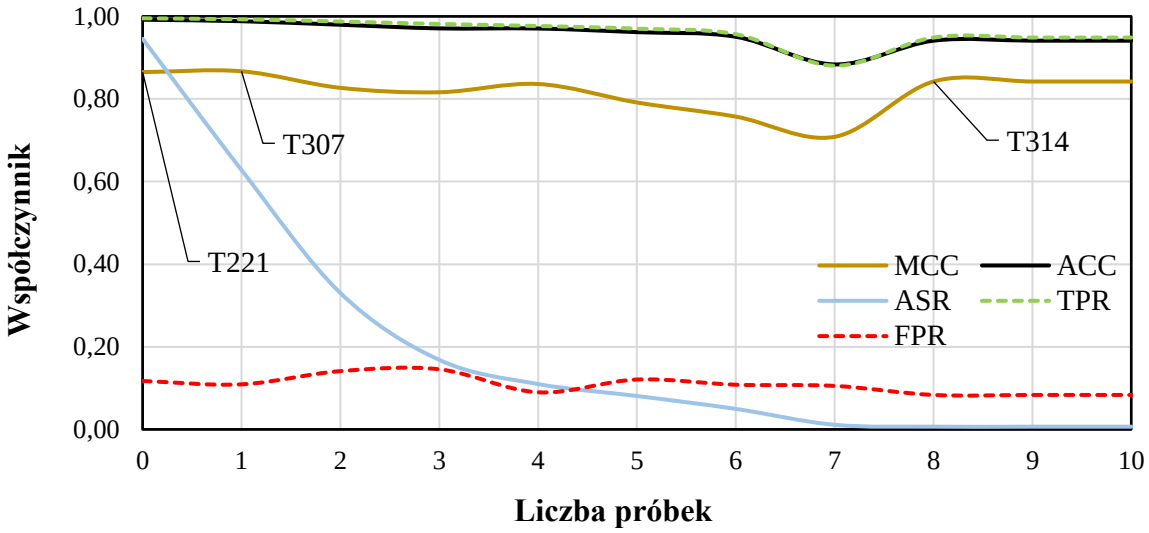
Zbadany został wpływ wymagania minimalnej liczby próbek z klasy pozytywnej (Rysunek 6.4) oraz negatywnej (Rysunek 6.5). Dla żądanego poziomu oceny próbek ze zbioru treningowego (ASR ≥ 80%), dla obu parametrów wyniki najlepsze (maksymalne: ACC, MCC, TPR; minimalne: FPR) zostały otrzymane dla punktu odniesienia (T221; Tabela 6.1), czyli gdy nie jest wyspecyfikowana liczba próbek. Szczególnie szybki spadek wskaźnika ASR jest widoczny dla klasy negatywnej (Ściek) – znaczna asymetria liczebności grup.

Testy T349 (Rysunek 6.4) charakteryzuje się największą wartością TPR = 99,20%, zaś T351 zmaksymalizowanym wskaźnikiem ACC = 99,23%. Natomiast w obu przypadkach współczynnik liczby sklasyfikowanych próbek w zbiorze treningowym zanotowano poniżej wymaganej granicy: ASR ≥ 80%. Gdy poddano analizie testy spełniają kryterium ASR dla T221 (punkt odniesienia) otrzymano najwyższe wartości dla wszystkich wskaźników.



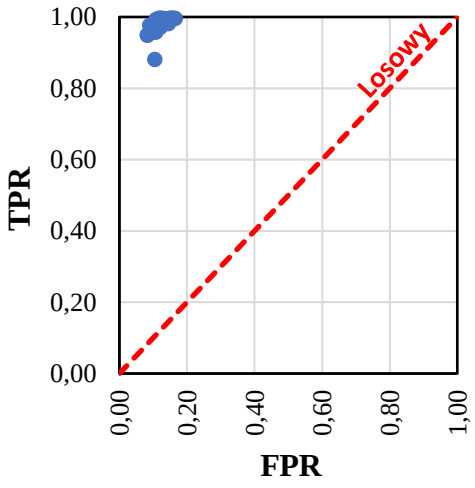
Rysunek 6.4. Profile wskaźników oceny klasyfikacji zarejestrowane podczas zmiany parametru minimalnej liczby próbek z klasy pozytywnej (Zawrócić).

W przypadku parametru definiującego wymaganą minimalną liczbę próbek z klasy negatywnej, kryterium ASR ≥ 80% nie jest osiągane już dla żądania co najmniej 1 pozycji (Rysunek 6.5) Świadczy to o asymetrii klas. Podczas pracy algorytmu dla pobranych pozycji z bazy produkcyjnej, istnieje niskie prawdopodobieństwo, że choć jedna z nich będzie miała przypisaną klasę negatywną (Ściek).



Rysunek 6.5. Profile wskaźników oceny klasyfikacji zarejestrowane podczas zmiany parametru minimalnej liczby próbek z klasy negatywnej (Ściek).

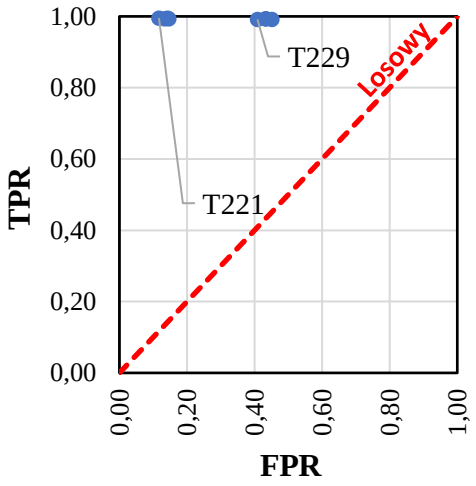
Rysunek 6.6 przedstawia układ współrzędnych TPR-FPR z naniesionymi punktami pochodzącymi z testów modyfikacji parametrów minimalnej liczby próbek dla klasy negatywnej oraz pozytywnej. Modyfikacja tych ustawień ma zauważalny wpływ na wartość wskaźników TPR (wskaźnik próbek prawdziwe pozytywnych) oraz FPR (współczynnik ocen fałszywie pozytywnych). Punkty są zgrupowane w lewej górnej części wykresu, oznacza to, że istnieje silna korelacja pomiędzy wynikami algorytmu, a rzeczywistymi etykietami pochodzącymi ze zbioru treningowego.



Rysunek 6.6. Układ współrzędnych TPR-FPR – wyniki ewaluacji zmiany parametrów wartości minimalnej wymaganej liczby próbek dla klasy pozytywnej i negatywnej.

6.2.4. Dyskretne parametry algorytmu

W tej grupie testów weryfikowane były łącznie parametry dyskretne, które nie charakteryzują bezpośrednio próbki. W Tabeli 6.3 zestawione zostały wszystkie testy wykonane w ramach tej grupy, łącznie z informacją o wykorzystaniu każdego z argumentów. Na Rysunku 6.7 widoczne są dwa zgrupowania punktów względem wskaźnika FPR. Znacznie słabszą jakość przypisania klasy negatywnej (Ściek) zaobserwowano w sytuacji, w której stosowano korekcję braku wystąpień (T224, T665, T229, T227) – wskaźnik próbek fałszywie pozytywnych (FPR) osiągnął wartość przekraczającą 40%. Ocena klasy pozytywnej (Zawrócić) dla każdego wariantu pozostała wysoka (TPR_{MIN} = 99,14%)



Rysunek 6.7. Układ współrzędnych TPR-FPR – wyniki ewaluacji zmiany parametrów dyskretnych, które nie charakteryzują próbki.

Tabela 6.3. Kombinatoryczna matryca testów parametrów dyskretnych, które nie charakteryzują próbki.

Numer testu	Korekcja wystąpień	Dyskretyzacja (odchylenie standardowe)	Tylko bez zaleceń
T221	Nie	Nie	Tak
T226	Nie	Tak	Tak
T228	Nie	Tak	Nie
T222	Nie	Nie	Nie
T224	Tak	Nie	Tak
T665	Tak	Tak	Tak
T229	Tak	Tak	Nie
T227	Tak	Nie	Nie

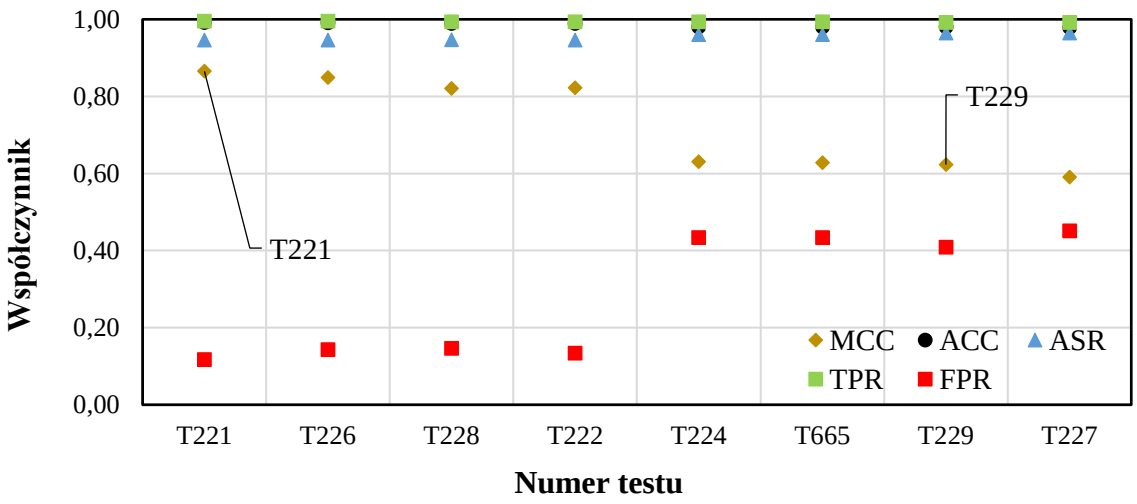
Na Rysunku 6.8 obserwuje się wysokie wartości wskaźników:

– dokładności (ACC_{MIN} = 97,85%; ACC_{MAX} = 99,20%),

– ocenionych próbek w zbiorze treningowym (ASR_{MIN} = 94,57%; ACC_{MAX} = 96,39%).

– ocena próbek prawdziwie pozytywnych (TPR_{MIN} = 99,14%; TPR_{MAX} = 99,54%).

Punkty serii ACC, ASR oraz TPR pokrywają się na Rysunku 6.8. Wraz z włączeniem korekcji wystąpień znacząco zwiększał się udział próbek oznaczonych fałszywie pozytywnie ($40{,}85\% \leq FPR \leq 45{,}11\%$), co odzwierciedlało się również w wynikach współczynnika korelacji (MCC), którego wartość spadała poniżej 0,65.



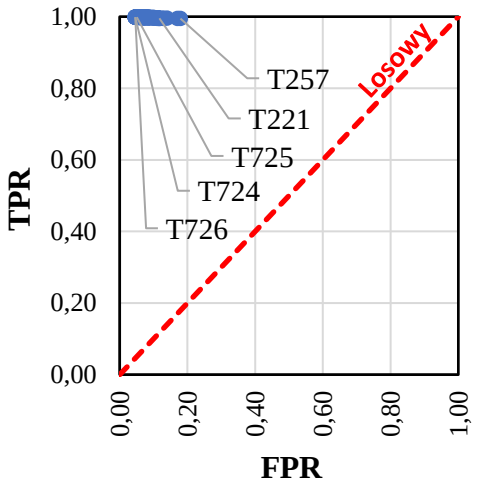
Rysunek 6.8. Profile wskaźników oceny klasyfikacji zarejestrowane podczas zmiany parametrów dyskretnych, które nie charakteryzują próbki.

6.3. Modyfikacje parametrów charakteryzujących próbkę

Zgodnie z planem testów parametry algorytmu odnoszące się do analiz fizykochemicznych były modyfikowane poprzez zmianę ich aktywności oraz przez inkrementację liczby poziomów dyskretnych. Dla wszystkich argumentów rzeczywistych zdefiniowano 90 wariantów testowych. Parametry dyskretne połączono w dwie grupy, które testowano w sposób kombinatoryczny, każda z nich zawierała 8 testów.

6.3.1. Analiza pH

Na Rysunku 6.9 zestawiono wszystkie punkty otrzymane w ramach modyfikacji tej pozycji ustawień. W Tabeli 6.4 zestawiono testy, w których kluczowe wskaźniki osiągnęły wartości maksymalne. Zauważalne jest to, że najlepsze wyniki otrzymano dla znacznie większej liczby poziomów dyskretnych (500 i więcej przedziałów), niż to było założone w punkcie odniesienia (5 przedziałów). Istotne jest to, że wyniki charakteryzujące się najniższą jakością klasyfikacji otrzymano dla testu (T257), w którym parametr pH nie był uwzględniany w klasyfikacji – oznaczony jest liczbą poziomów dyskretnych równą 0. Istnieje więc pozytywna korelacja pomiędzy liczbą poziomów dyskretnych, a jakością klasyfikacji. Obserwowany jest silniejszy wpływ na klasę negatywną (FPR maleje wraz ze wzrostem liczby poziomów dyskretnych, FPR_{MIN} = 4,59%; FPR_{MAX} = 17,94%) niż na klasę pozytywną (TPR pozostaje na stabilnym poziomie dla wszystkich wariantów).

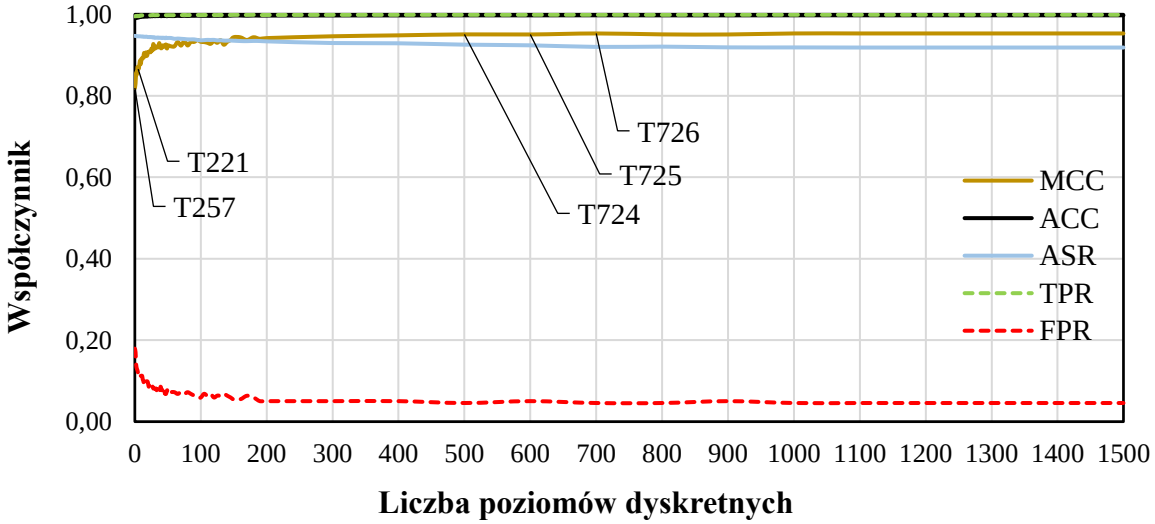


Rysunek 6.9. Układ współrzędnych TPR-FPR – wyniki ewaluacji modyfikacji parametru pH.

Tabela 6.4. Zestawienie testów, w których współczynnik osiągnął maksimum – modyfikacja parametru pH.

Numer testu	Zmaksymalizowany współczynnik	Wartość nastawy
T726	ACC	700
T726	MCC	700
T257	ASR	0
T725	TPR	600
T724	TNR = (1 - FPR)	500
T221		5

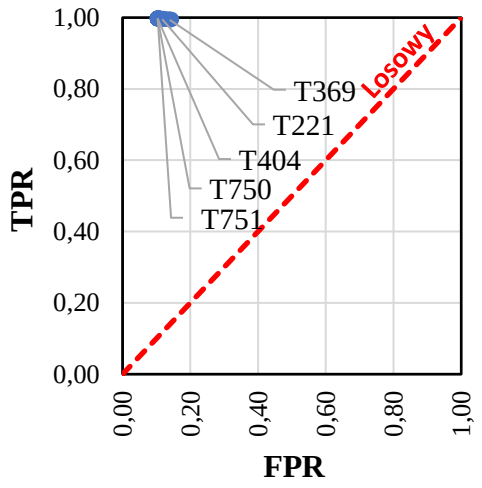
Na Rysunku 6.10 przedstawiono charakterystykę wpływu modyfikacji parametru analizy pH na wskaźniki jakości algorytmu. Wartości metryk stabilizowały się po przekroczeniu 100 poziomów dyskretnych. Zauważalny jest systematyczny spadek liczby próbek, którym przyznano jakąkolwiek klasę (ASR_{MIN} = 91,81%; ASR_{MAX} = 94,68%). Krzywe TPR oraz ACC nałożyły się na siebie. W teście 100 poziomów dyskretnych 29 próbek zostało błędnie zaklasyfikowane, zaś dla najwyższej wartości dokładności (ACC_{MAX} = 99,72%) liczba błędny klasyfikacji wyniosła 20 sztuk. Wyniki dokładności były na wysokim poziomie dla każdego wariantu (ACC_{MIN} = 98,94%). Współczynnik korelacji MCC rósł wraz z dokładnością, osiągnął swoje maksimum (MCC_{MAX} = 0,9527) dla tego samego wariantu – T726.



Rysunek 6.10. Profile wskaźników oceny klasyfikacji zarejestrowane podczas zmiany parametru odpowiadającego analizie pH.

6.3.2. Analiza gęstości

Parametr odpowiadający wynikom pomiarów gęstości został poddany testom. Brak uwzględnienia tych analiz w klasyfikacji również skutkował najniższymi ocenami jakościowymi (T369). Zanotowano korelację pomiędzy jakością klasyfikacji, a liczbą przedziałów dyskretnych. Na Rysunku 6.11 przedstawiono punkty wszystkich testów. Modyfikacja parametru gęstości miała wpływ na wskaźnik liczby próbek fałszywie pozytywnych (FPR_{MIN} = 10,36%; FPR_{MAX} = 14,41%; zmiana odpowiadała 9 próbkom). W zestawieniu testów (Tabela 6.5) zauważalne są znacznie mniejsze wartości nastaw (liczby przedziałów) niż to miało miejsce w przypadku pH.

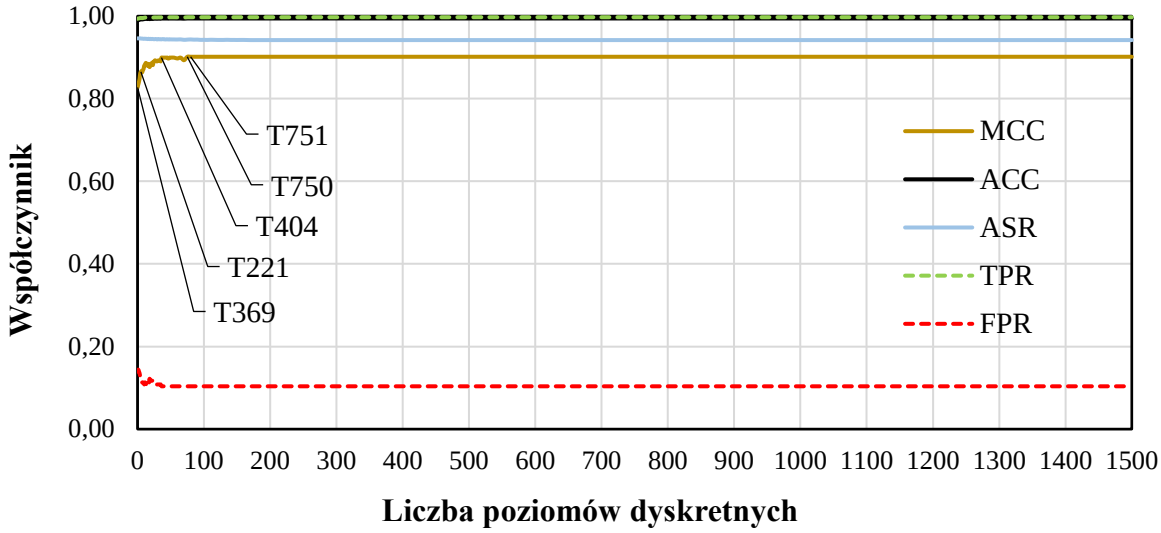


Rysunek 6.11. Układ współrzędnych TPR-FPR – wyniki ewaluacji modyfikacji parametru gęstości.

Tabela 6.5. Zestawienie testów, w których współczynnik osiągnął maksimum – modyfikacja parametru gęstości.

Numer testu	Zmaksymalizowany współczynnik	Wartość nastawy
T751	ACC	80
T750	MCC	75
T369	ASR	0
T751	TPR	80
T404	TNR = (1 - FPR)	36
T221		5

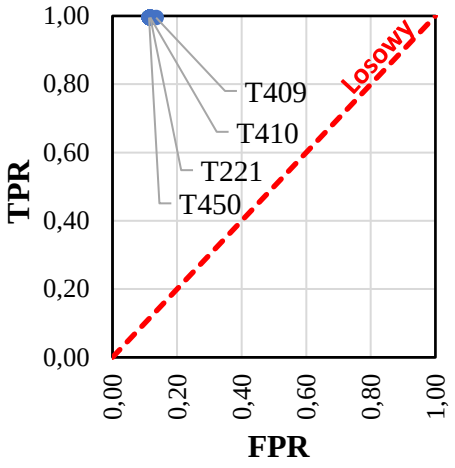
Stabilizacja wartości współczynników oceny jakościowej narzędzia klasyfikacji miała miejsce po przekroczeniu 36 przedziałów (Rysunek 6.12). Dokładność (ACC) zawarła się w przedziale pomiędzy 99,43% a 98,95%. Tak więc, wariant o najniższej dokładności (T369) zaklasyfikował błędnie 9 próbek więcej, niż test o najwyższej wartości ACC (T571). Pomijalny był wpływ modyfikacji parametru na zdolność przypisania klasy próbkom w zbiorze treningowym (ASR_{MIN} = 94,19%; ASR_{MAX} = 94,62%). Współczynnik korelacji osiągnął najwyższą wartość w wariancie T750 (MCC_{MAX} = 0,9016). Maksymalne wartości metryk oceny algorytmu otrzymano dla wariantów z liczbą przedziałów dyskretnych mniejszą od 100.



Rysunek 6.12. Profile wskaźników oceny klasyfikacji zarejestrowane podczas zmiany parametru odpowiadającego analizie gęstości.

6.3.3. Analiza lepkości

Zmiana wartości parametru algorytmu powiązanego z analizą lepkości pokazuje jego znikomy wpływ na jakość działania klasyfikacji. Zgodnie z matrycą testów dla parametrów rzeczywistych przeprowadzano 90 testów – wszystkie punkt naniesiono na Rysunku 6.13. Widoczne jest wzajemne nakładanie się punktów, co oznacza, że zarówno parametr TPR jak i FPR nie wykazują silnej korelacji z liczbą poziomów dyskretnych wyników analiz lepkości. W Tabeli 6.6 zestawione zostały testy, dla których współczynniki algorytmu osiągnęły maksymalne wartości oraz ewaluowaną liczbę poziomów dyskretnych. Warianty o niskich liczbach przedziałów dyskretnych (mniejszych od 50) charakteryzowały się najlepszą jakością klasyfikacji próbek ze zbioru treningowego.

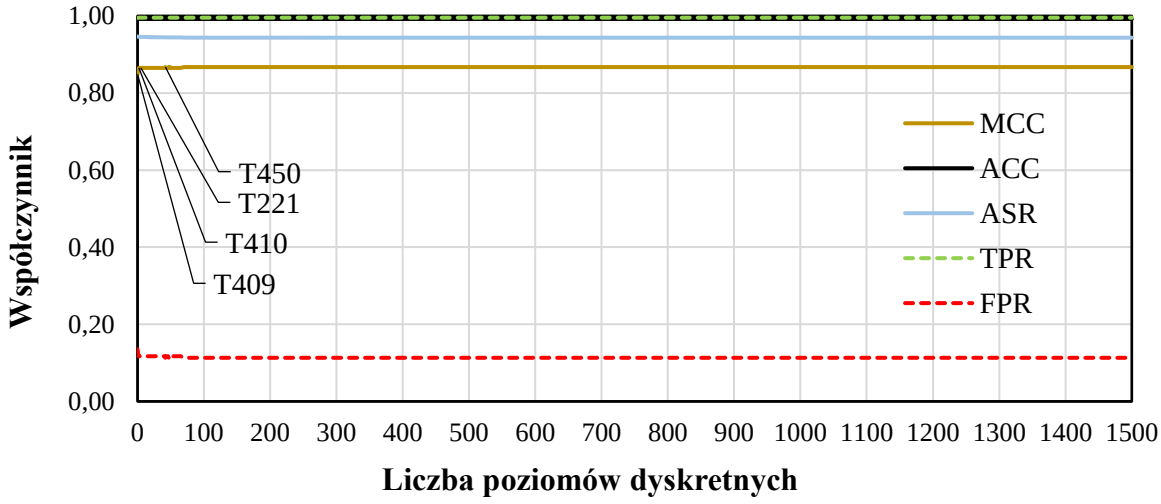


Rysunek 6.13. Układ współrzędnych TPR-FPR – wyniki ewaluacji modyfikacji parametru lepkości.

Tabela 6.6. Zestawienie testów, w których współczynnik osiągnął maksimum – modyfikacja parametru lepkości.

Numer testu	Zmaksymalizowany współczynnik	Wartość nastawy
T450	ACC	42
T450	MCC	42
T409	ASR	0
T410	TPR	2
T450	TNR = (1 - FPR)	42
T221		5

Profile wskaźników przedstawiono na Rysunku 6.14 Widoczne są stabilne poziomy otrzymane dla wszystkich metryk. Dokładność (ACC) zmieniała się od wartości 99,10% do 99,21%. Najniższą dokładność otrzymano dla testu, w którym parametr był wyłączony (T409), a najwyższą dla wariantu, w którym zdefiniowano 42 poziomy dyskretne (T450), ten ostatni przypisał prawidłową klasę 9 próbkom więcej. Dla liczby poziomów dyskretnych mniejszej niż 25 nie obserwuje się słabszej jakości klasyfikacji (niższych wartości MCC oraz wyższych wartości FPR), jak to zaobserwowano w przypadku poprzednich parametrów rzeczywistych. Dla wartości nastawy równej 2 przedziałom dyskretnym (T410) zaobserwowano spadek FPR o 1,8 punktu procentowego względem najwyższej wartości (T409; 13,51%), kolejne optymalizacje pozwoliły na dalsze obniżenie wartości jedynie o 0,4 pp. (T450).

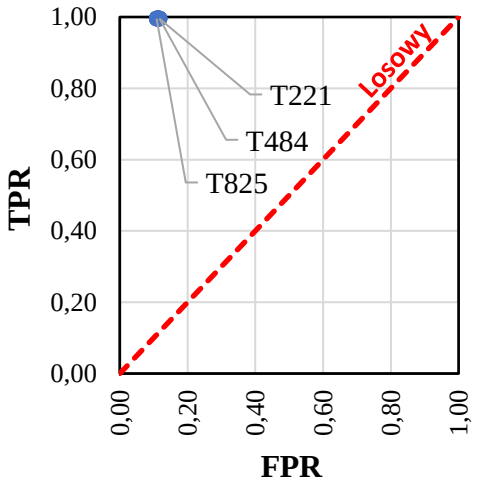


Rysunek 6.14. Profile wskaźników oceny klasyfikacji zarejestrowane podczas zmiany parametru odpowiadającego analizie lepkości.

Na Rysunku 6.14 obserwuje się pomijalne zmiany w wartości wskaźnika ocenionych próbek ze zbioru treningowego (ASR_{MIN} = 94,35%; ASR_{MAX} = 94,67%). W związku z tym, że w teście T450 próbki z klasy negatywnej (Ściek) zostały w większej liczbie poprawnie zaklasyfikowane, to współczynniki korelacji (zbalansowana metryka) zmienił się zauważalnie z MCC_{MIN} = 0,8488 do MCC_{MAX} = 0,8672.

6.3.4. Analiza stężenia procentowego nadtlenku wodoru

Nadtlenek wodoru jest substancją aktywną używaną w części produktów. Wartość oznaczonego stężenia jest krytycznym parametrem decydującym o przydatności produktu. Podczas testowania modyfikacji tego argumentu algorytmu nie zaobserwowano znaczących zmian wskaźników oceny jakości klasyfikacji (Rysunek 6.15). Otrzymana minimalna dokładność wynosiła ACC_{MIN} = 99,19%, a maksymalna ACC_{MAX} = 99,23%. Oznacza to, że wariant charakteryzujący się największą dokładnością sklasyfikował poprawnie 3 próbki więcej. Odpowiednio współczynnik korelacji MCC zmienił się zauważalnie (MCC_{MIN} = 0,8651; MCC_{MAX} = 0,8717), wiązało się to z lepszą klasyfikacją próbek negatywnych (FP zmalało o 2). Tabela 6.7 zawiera zestawienie testów, dla których wskaźniki osiągnęły maksimum. Warte odnotowania jest to, że dla testu T825 (600 przedziałów dyskretnych) wszystkie kluczowe metryki osiągnęły maksimum.

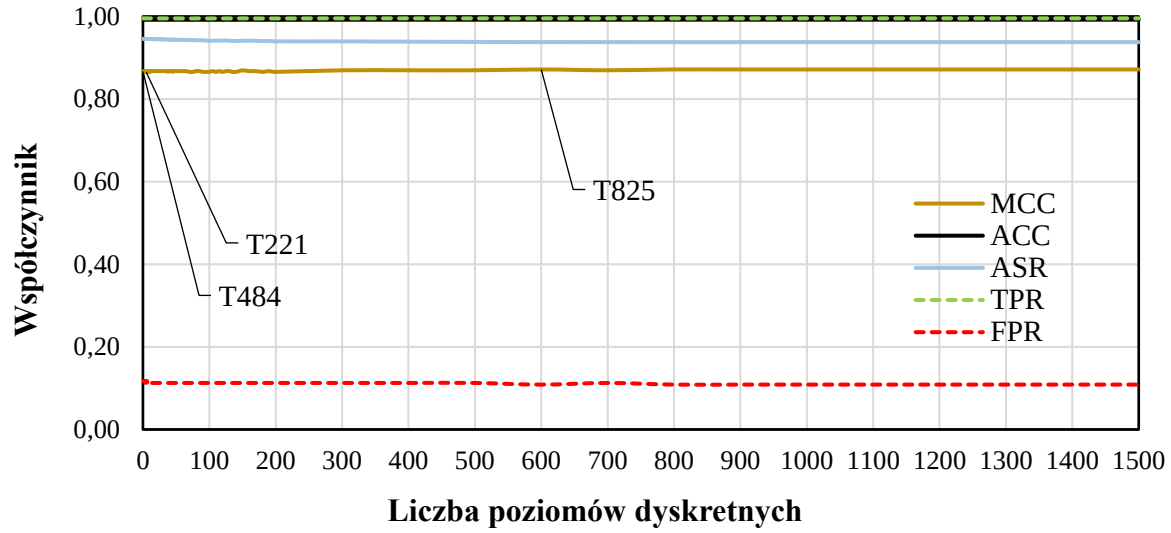


Rysunek 6.15. Układ współrzędnych TPR-FPR – wyniki ewaluacji modyfikacji parametru stężenia proc. nadtlenku wodoru.

Tabela 6.7. Zestawienie testów, w których współczynnik osiągnął maksimum – zmiana parametru stężenia proc. nadtlenku wodoru.

Numer testu	Zmaksymalizowany współczynnik	Wartość nastawy
T825	ACC	600
T825	MCC	600
T484	ASR	0
T484	TPR	0
T825	TNR = (1 - FPR)	600
T221		5

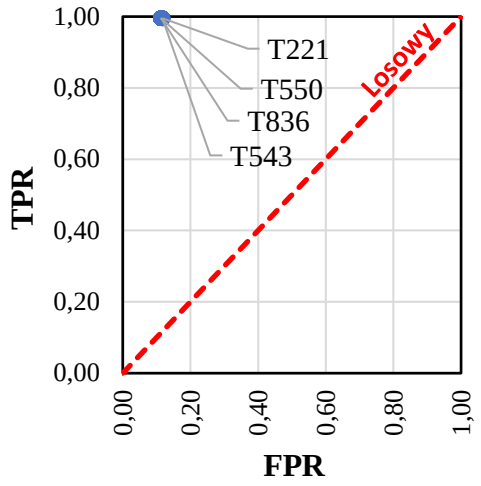
Profile wskaźników oceny algorytmu przedstawione na Rysunku 6.16 charakteryzują się stabilną wartością dla wszystkich wariantów modyfikacji. Parametr oceniający jaka część próbek ze zbioru treningowego została oceniona osiągną najniższą wartość ASR_{MIN} = 93,78% (T484). W teście T484 zaobserwowano ASR_{MAX} = 94,58%.



Rysunek 6.16 Profile wskaźników oceny klasyfikacji zarejestrowane podczas zmiany parametru odpowiadającego analizie stężenia procentowego nadtlenku wodoru.

6.3.5. Analiza stężenia wolnego chloru

W produktach dezynfekujących wykorzystuje się związki chloru (podchloryn sodu). Rutynowo wykonuje się analizę miareczkową tzw. wolnego chloru, czyli takiego, który jest zdolny do działania i może natychmiast spowodować utlenianie niepożądanych substancji. Jest to analiza krytyczna dla oceny produktu, ponieważ wykazuje ona zgodność ze specyfikacją zawartości substancji czynnej. Modyfikacja parametru algorytmu odpowiadająca analizie chloru wykazała niską wpływ na dokładność oceny klasyfikacji. Nie zaobserwowano istotnej zmiany parametrów TPR oraz FPR (Rysunek 6.17). Dla testu, z wyłączonym parametrem otrzymano najniższą wartość ACC_{MIN} = 99,17%. Najwyższa dokładność wyniosła ACC_{MAX} = 99,26%. Wynika z tego, że optymalna liczba poziomów dyskretnych poprawiła skuteczność klasyfikacji o 7 poprawnych oznaczeń. Zauważalne jest to, że dla jednego wariantu maksimum osiągnęły 3 wskaźniki ACC, MCC oraz TPR (Tabela 6.8).

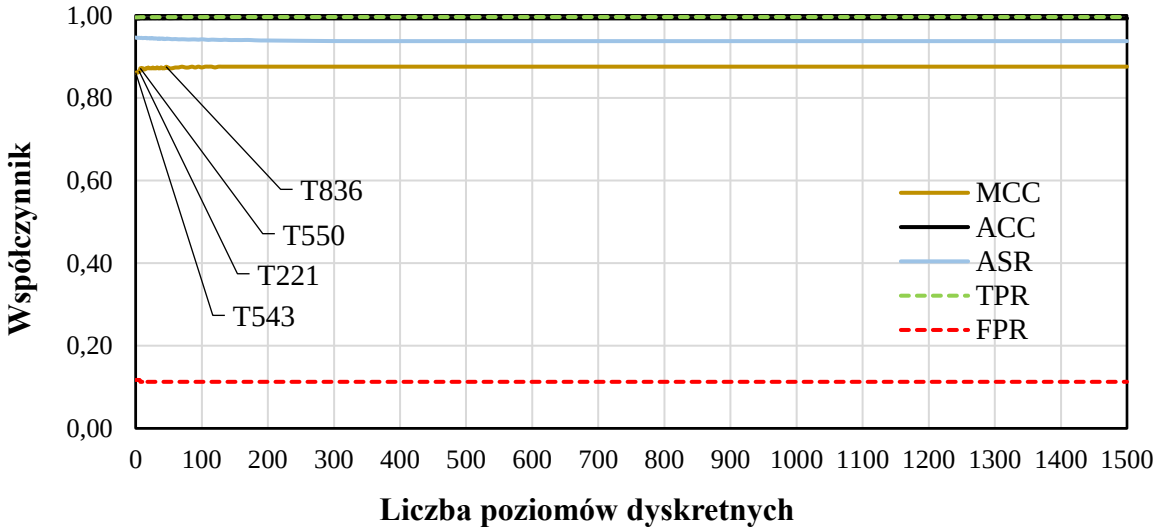


Rysunek 6.17. Układ współrzędnych TPR-FPR – wyniki ewaluacji modyfikacji parametru stężenia proc. wolnego chloru.

Tabela 6.8. Zestawienie testów, w których współczynnik osiągnął maksimum – zmiana parametru stężenia proc. wolnego chloru.

Numer testu	Zmaksymalizowany współczynnik	Wartość nastawy
T836	ACC	46
T836	MCC	46
T543	ASR	0
T836	TPR	46
T550	TNR = (1 - FPR)	7
T221		5

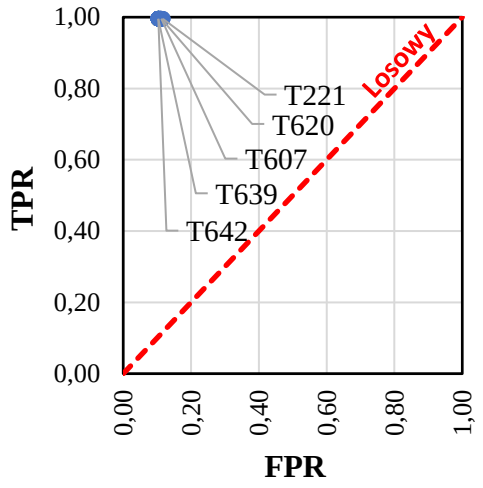
Na Rysunku 6.18 obserwuje się stały poziom wartości wskaźników w domenie liczby przedziałów dyskretnych. Wszystkie maksima metryk osiągnięto dla wartości liczby poziomów mniejszej od 50. Wskaźnik korelacji zmieniał się od MCC_{MIN} = 0,8612 do MCC_{MAX} = 0,8757. Również zaobserwowano niską zależność pomiędzy modyfikacją parametru, a możliwością przypisania klasy przez algorytm (ASR_{MIN} = 93,76%; ASR_{MAX} = 94,63%).



Rysunek 6.18 Profile wskaźników oceny klasyfikacji zarejestrowane podczas zmiany parametru odpowiadającego analizie stężenia procentowego wolnego chloru.

6.3.6. Analiza suchej pozostałości

Analiza suchej pozostałości to wartość masy pozostałej po odparowaniu rozpuszczalników w temperaturze 105°C. Wynikiem analizy jest procentowa wartość stanowiąca stosunek masy pozostałej po suszeniu do masy pierwotnej próbki. Również w tym przypadku nie jest obserwowana silna zależność pomiędzy liczbą poziomów dyskretnych, a wskaźnikami TPR oraz FPR (Rysunek 6.19). W Tabeli 6.9 z testami, dla których kluczowe współczynniki osiągnęły wartości maksymalne, obserwuje się, że dla niskiej liczby przedziałów osiągane są maksima – najwyższa wartość nastawy to 1500 (Tabela 5.4).

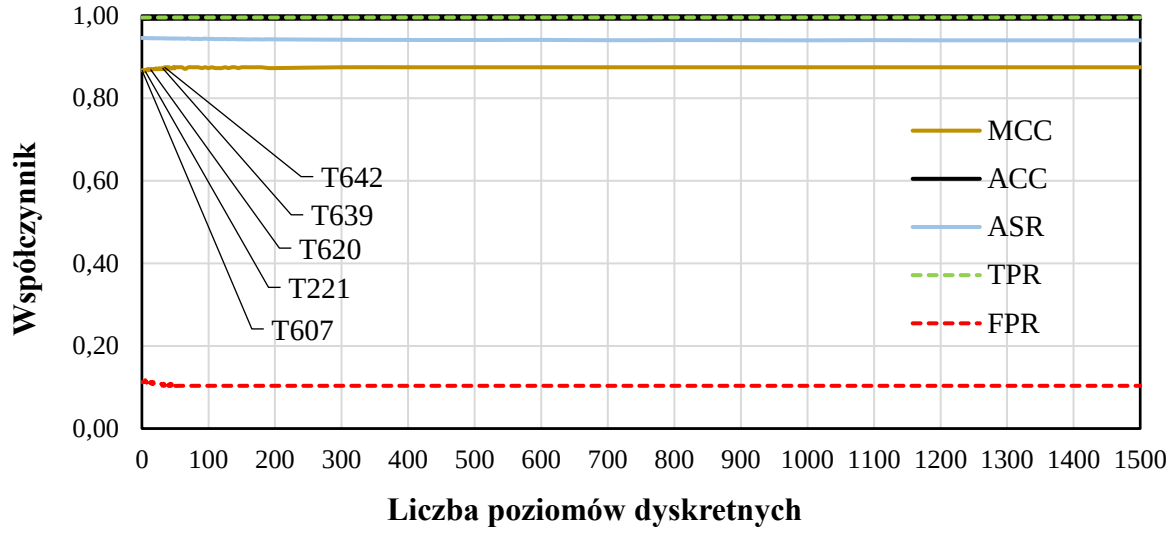


Rysunek 6.19. Układ współrzędnych TPR-FPR – wyniki ewaluacji modyfikacji parametru analizy suchej pozostałości.

Tabela 6.9. Zestawienie testów, w których współczynnik osiągnął maksimum – zmiana parametru analizy suchej pozostałości.

Numer testu	Zmaksymalizowany współczynnik	Wartość nastawy
T642	ACC	34
T642	MCC	34
T607	ASR	0
T620	TPR	13
T639	TNR = (1 - FPR)	31
T221		5

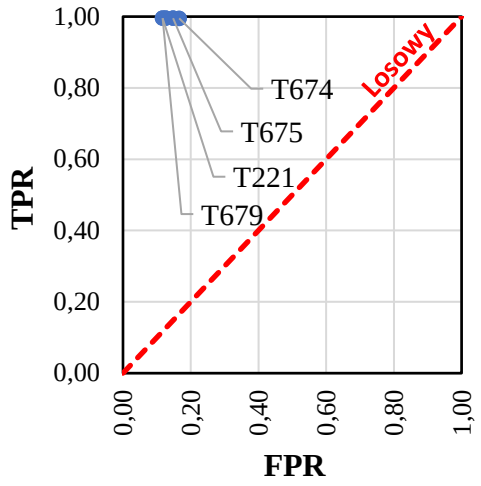
W przedstawionych profilach wskaźników (Rysunek 6.20), obserwowany jest spadek wartości FPR. Względem najniższej wartość FPR równej 11,26% (T607) optymalizacje pozwoliły o obniżenie wartości współczynnika o 1,35 punktu procentowego. Wersja testu o najmniejszej dokładności (ACC_{MIN} = 99,20%) oceniła błędnie 58 próbek (T612), zaś ta o najwyższej ocenie (ACC_{MAX} = 99,25%) 55 pozycji (T642). Na Rysunku 6.20 przedstawiono profile współczynników. Dla niskich wartości nastaw (mniejszych od 15) obserwowano spadek wartości FPR oraz wzrost współczynnika korelacji (MCC_{MIN} = 0,8651; MCC_{MAX} = 0,8749). Zmiany w wartościach bezwzględnych próbek TP oraz FP są pomijalne – różnica pomiędzy minimum i maksimum nie przekracza 2.



Rysunek 6.20. Profile wskaźników oceny klasyfikacji zarejestrowane podczas zmiany parametru odpowiadającego analizie suchej pozostałości.

6.3.7. Parametry dyskretne – cechy partii produktu

Każdej próbce przypisano parametry, które nie stanowiły jej bezpośrednich parametrów, tylko charakteryzowały partię półproduktu. Zgodnie z Tabelą 5.1 były to: rodzaj (proces, z którego pochodzi wyrób), typ (definicja zgodności), instalacja (miejsce wytworzenia). Parametry te testowano kombinatorycznie w ramach jednej grupy (Tabela 6.10). Na Rysunku 6.21 obserwuje się wpływ nastaw na wartość współczynnika FPR. Zaobserwowano, że dla wszystkich testów, które nie uwzględniały instalacji (T675, T678, T674, T676) otrzymano najniższe oceny jakości algorytmu (niższe wartości MCC, ACC, TPR oraz wyższe wartości FPR) – Rysunek 6.22. Wskaźnik próbek fałszywie pozytywnych zmieniał się od FPR_{MIN} = 11,71% (T679) do FPR_{MAX} = 16,59% (T674). Odzwierciedliło się to również w wartościach współczynnika korelacji (MCC_{MIN} = 0,8328; MCC_{MAX} = 0,8651). Nie zanotowano znaczącego wpływu na ogólną dokładność klasyfikacji (ACC_{MIN} = 99,03%; ACC_{MAX} = 99,20%) oraz na wskaźnik liczby próbek prawdziwie pozytywnych (TPR_{MIN} = 99,51%; TPR_{MAX} = 99,55%).

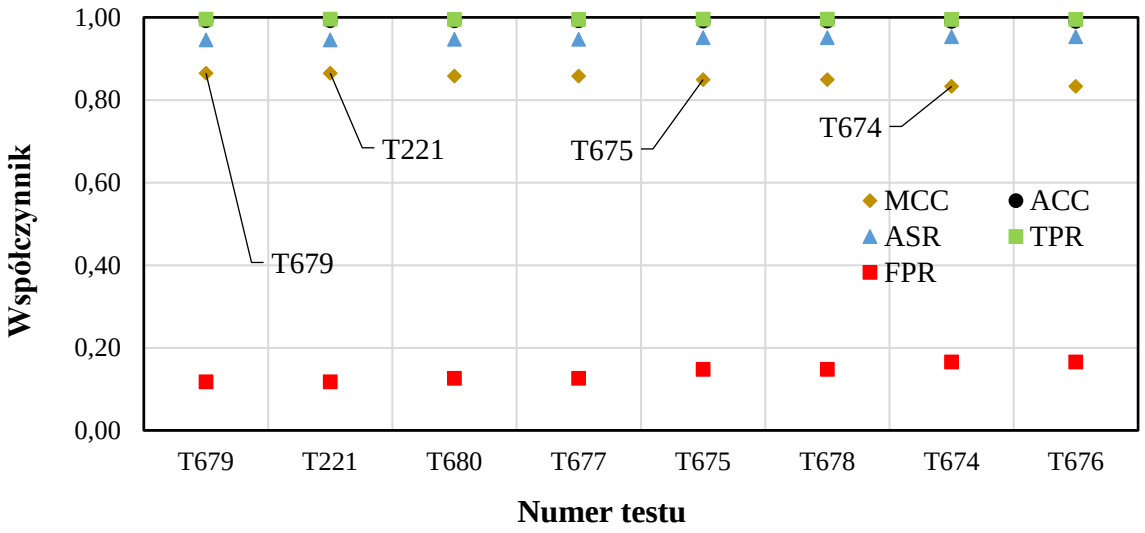


Rysunek 6.21. Układ współrzędnych TPR-FPR – wyniki ewaluacji modyfikacji parametrów dyskretnych, które stanowiły cechy partii półproduktu.

Tabela 6.10. Kombinatoryczna matryca testów parametrów dyskretnych, które stanowiły cechy partii półproduktu.

Numer testu	Rodzaj	Typ	Instalacja
T679	Tak	Nie	Tak
T221	Tak	Tak	Tak
T680	Nie	Tak	Tak
T677	Nie	Nie	Tak
T675	Tak	Nie	Nie
T678	Tak	Tak	Nie
T674	Nie	Nie	Nie
T676	Nie	Tak	Nie

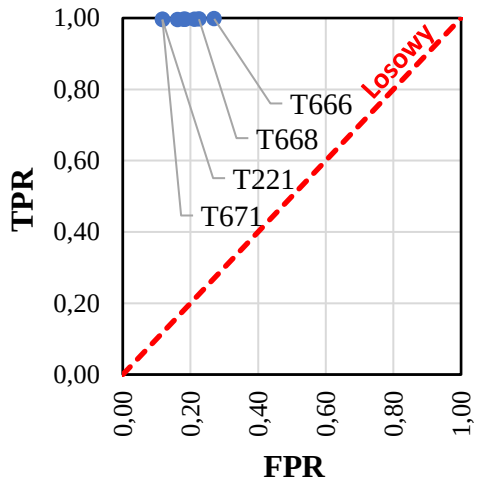
Dokładnie takie same wartości wszystkich metryk otrzymano dla ogólnego punktu odniesienia (T221) oraz dla wariantu testu, który nie uwzględniał typu półproduktu (T679).



Rysunek 6.22. Profile wskaźników oceny klasyfikacji zarejestrowane podczas zmiany parametrów dyskretnych, które stanowiły cechy partii półproduktu.

6.3.8. Parametry dyskretne – porównanie produktu ze wzorcem

Laboratorium kontroli jakości w ramach analizy fizykochemicznych weryfikowało, czy produkt jest zgodny ze wzorcem. Analizie poddawane były 3 parametry: kolor, zapach i wygląd. Każdy parametr mógł zostać oznaczony jako zgodny lub niezgodny ze standardem. W ramach badań optymalizacyjnych algorytmu klasyfikacji parametry te były testowane łącznie w sposób kombinatoryczny (Tabela 6.11). Wyniki wskaźników TPR oraz FPR przedstawiono na Rysunku 6.23. Widoczny jest znikomy wpływ na wartość wskaźnika próbek prawdziwie pozytywnych (TPR_{MIN} = 99,54%; TPR_{MAX} = 99,76%). Natomiast metryka oceniająca liczbę klasyfikacji fałszywie pozytywnych zmienia się znacząco wraz z wersjami testów (FPR_{MIN} = 11,71%; FPR_{MAX} = 27,03%). Najlepsze odpowiedzi uzyskano dla testu, który nie uwzględniał jedynie zapachu (T671) oraz dla wariantu uwzględniającego wszystkie analizy (T221 – punkt odniesienia).

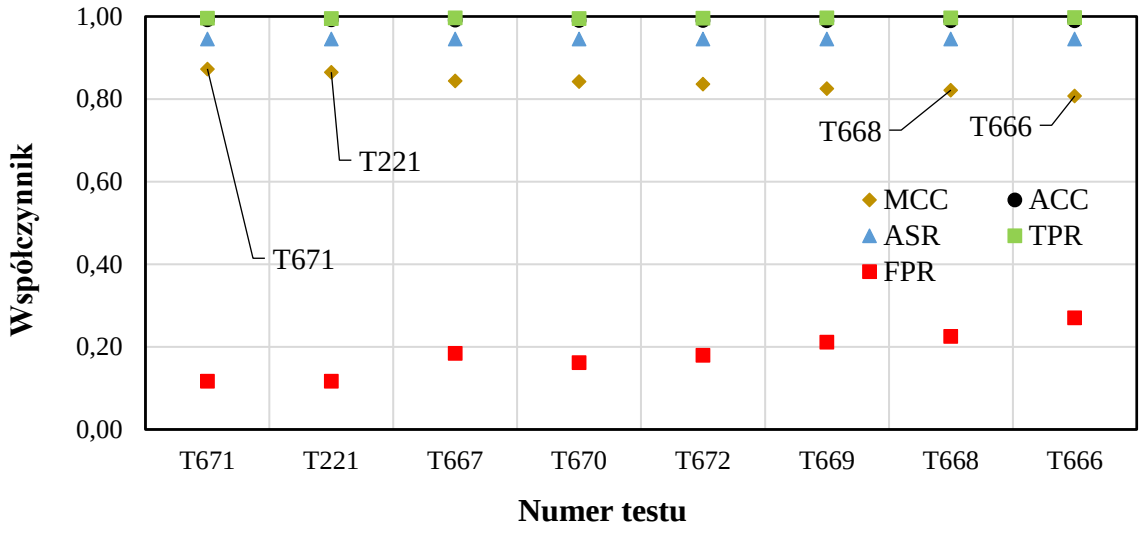


Rysunek 6.23. Układ współrzędnych TPR-FPR – wyniki ewaluacji modyfikacji parametrów dyskretnych, które odpowiadały analizom sensorycznym półproduktu.

Tabela 6.11. Kombinatoryczna matryca testów parametrów dyskretnych, które odpowiadały analizom sensorycznym.

Numer testu	Kolor	Wygląd	Zapach
T671	Tak	Tak	Nie
T221	Tak	Tak	Tak
T667	Tak	Nie	Nie
T670	Tak	Nie	Tak
T672	Nie	Tak	Tak
T669	Nie	Tak	Nie
T668	Nie	Nie	Tak
T666	Nie	Nie	Nie

Na profilach wskaźników przestawionych na Rysunku 6.24 obserwuje się również silnie zmieniającą się wartość współczynnika korelacji (MCC_{MIN} = 0,8076; MCC_{MAX} = 0,8731). Dokładność osiągnęła swoje maksimum dla testu T671 (ACC_{MAX} = 99,25%), zaś najmniejszą jej wartość uzyskano dla testu, który nie uwzględniał żadnego parametru (ACC_{MIN} = 98,95%).



Rysunek 6.24. Profile wskaźników oceny klasyfikacji zarejestrowane podczas zmiany parametrów dyskretnych, które odpowiadały analizom sensorycznym półproduktu.

6.4. Porównanie testów modyfikacji jednego parametru

Grupy testów opisane w rozdziałach 6.2 oraz 6.3 polegały na iteracyjnej zmianie badanego parametru (grupy kombinatorycznej), podczas gdy pozostałe argumenty nie ulegały modyfikacjom – przyjmowały wartość ustaloną jako punk odniesienia (opisany w Rozdziale 6.1). Kryterium sukcesu zostało oparte o wskaźnik dokładności (Rozdział 5.2.6). W ramach przeprowadzonych testów zaobserwowano jednak, że zbiór danych jest niewystarczająco zbalansowany. Tak więc niedokładność w klasie negatywnej jest maskowana przez wysoką dokładność liczniejszej klasy pozytywnej. Postanowiono więc do dalszych badań optymalizacyjnych wytypować dwóch kandydatów z każdej grupy testów. Zgodnie z pierwotnymi założeniami, pierwszym z nich jest ten wariant, który osiągnął maksymalną wartość dokładności (ACC_{MAX}). Drugim testem dołączonym do matrycy badań był ten, który osiągnął maksymalną wartość współczynnika korelacji Matthewsa (MCC_{MAX}). Ten ostatni jest miarą zbalansowaną, a więc jest czuły na dokładność w każdej z klas – niweluje problem

wynikający z różnej ich liczności. Poniżej w Tabeli 6.12 przedstawiono wartości wskaźników oceny algorytmu dla przyjętego punktu odniesienia (T221).

Tabela 6.12. Wartości wskaźników oceny ewaluacji dla ogólnego punktu odniesienia (T221).

Testowany wariant	Numer testu	MCC	ACC	ASR	TP	TN	FP	FN
Ogólny punkt odniesienia	T221	0,8651	99,20%	94,57%	7090	196	26	33

W Tabeli 6.13 zestawione zostały badane nastawy algorytmu. W ramach każdej kategorii przedstawiono test (kolumna druga), który uzyskał w ramach iteracyjnych badań najwyższą wartość dokładności (ACC) – w czwartej kolumnie przytoczono wartość liczbową. Nie jest to najwyższa wartość dla populacji wszystkich przeprowadzonych testów, tylko lokalne maksimum dla badanej kategorii. W pozostałych kolumnach przytoczono wartości współczynnika korelacji (MCC) oraz zdolności przypisania klasy (ASR). Przestawione zostały również wartości kategorii zgodnie z tablicą pomyłek (TP, TN, FP, FN). Przygotowana została również analogiczna Tabela 6.14 w ramach rozbudowania zakresu badań o wskaźnik zbalansowany. Zawarto w niej testy, które dla danej kategorii osiągnęły lokalne maksimum współczynnika korelacji Matthewsa (MCC). Porównanie wartości z obu tabel pokazuje, że jedynie parametr odpowiadający analizie gęstości nie uzyskał wartości maksymalnych ACC oraz MCC w tym samym wariancie testu. Okazało się więc, że pomimo dużego niezbalansowania liczebności klas, kryterium sukcesu oparte na dokładności klasyfikacji uznać można za prawidłowo zdefiniowane. Dane w Tabeli 6.13 uszeregowane są malejąco względem ACC, zaś w Tabeli 6.14 względem MCC. Zauważalny jest znacznie większy rozrzut wyników współczynnika korelacji (MCC_{MIN} = 0,8651; MCC_{MAX} = 0,9527) niż dokładności (ACC_{MIN} = 99,72%; ACC_{MAX} = 99,20%). Ze względów prowadzenia badań optymalizacyjnych obie metryki można uznać za tak samo wartościowe. Dodatkowo jednak MCC wskazuje, które parametry miały największy wpływ na poprawę skuteczności klasyfikacji.

Tabela 6.13. Zestawienie testów wyników ewaluacji parametrów, dla których wartość współczynnika dokładności (ACC) osiągnęła maksimum.

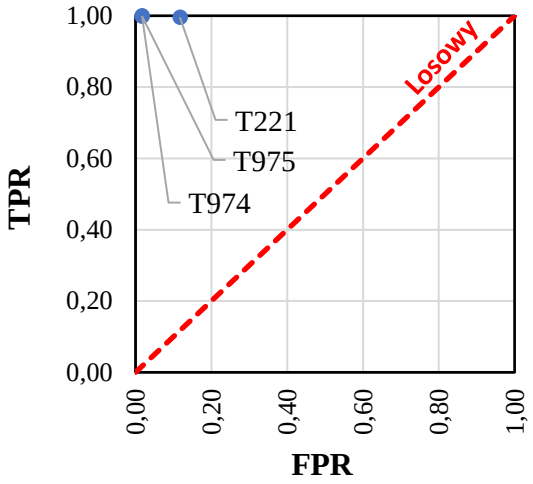
Testowany parametr	Numer testu	MCC	ACC	ASR	TP	TN	FP	FN
Analiza pH	T726	0,9527	99,72%	91,97%	6915	208	10	10
Analiza gęstości	T751	0,9016	99,43%	94,35%	7087	199	23	19
Analiza stężenia wolnego chloru	T836	0,8757	99,26%	94,30%	7073	197	25	29
Parametry dyskretne – porównanie ze wzorcem produktu	T671	0,8731	99,25%	94,57%	7094	196	26	29
Analiza suchej pozostałości	T642	0,8749	99,25%	94,52%	7087	199	23	32
Maksymalna ogólna liczba próbek uwzględniana do obliczeń	T238	0,8716	99,23%	95,75%	7180	200	23	34
Analiza stężenia procentowego nadtlenku wodoru	T825	0,8717	99,23%	93,79%	7032	197	24	32
Analiza lepkości	T450	0,8672	99,21%	94,44%	7081	196	25	33
Wymagana minimalna ogólna liczba próbek	T894	0,8718	99,20%	95,38%	7141	208	26	33
Wymagana minimalna liczba próbek z klasy pozytywnej (Zawrócić)	T221	0,8651	99,20%	94,57%	7090	196	26	33
Dyskretne parametry algorytmu	T221	0,8651	99,20%	94,57%	7090	196	26	33
Parametry dyskretne – cechy partii produktu	T679	0,8651	99,20%	94,57%	7090	196	26	33
Wymagana minimalna liczba próbek z klasy pozytywnej (Ściek)	T221	0,8651	99,20%	94,57%	7090	196	26	33

Tabela 6.14. Zestawienie testów wyników ewaluacji parametrów, dla których wartość współczynnika korelacji Matthewsa (MCC) osiągnęła maksimum.

Testowany parametr	Numer testu	MCC	ACC	ASR	TP	TN	FP	FN
Analiza pH	T726	0,9527	99,72%	91,97%	6915	208	10	10
Analiza gęstości	T750	0,9016	99,43%	94,31%	7084	199	23	19
Analiza stężenia wolnego chloru	T836	0,8757	99,26%	94,30%	7073	197	25	29
Analiza suchej pozostałości	T642	0,8749	99,25%	94,52%	7087	199	23	32
Parametry dyskretne – porównanie ze wzorcem produktu	T671	0,8731	99,25%	94,57%	7094	196	26	29
Wymagana minimalna ogólna liczba próbek	T894	0,8718	99,20%	95,38%	7141	208	26	33
Analiza stężenia procentowego nadtlenku wodoru	T825	0,8717	99,23%	93,79%	7032	197	24	32
Maksymalna ogólna liczba próbek uwzględniana do obliczeń	T238	0,8716	99,23%	95,75%	7180	200	23	34
Analiza lepkości	T450	0,8672	99,21%	94,44%	7081	196	25	33
Wymagana minimalna liczba próbek z klasy pozytywnej (Ściek)	T221	0,8651	99,20%	94,57%	7090	196	26	33
Wymagana minimalna liczba próbek z klasy pozytywnej (Zawrócić)	T221	0,8651	99,20%	94,57%	7090	196	26	33
Dyskretne parametry algorytmu	T221	0,8651	99,20%	94,57%	7090	196	26	33
Parametry dyskretne – cechy partii produktu	T679	0,8651	99,20%	94,57%	7090	196	26	33

6.5. Jednoczesna modyfikacja wszystkich parametrów

Zgodnie z planem testów, po fazie optymalizacji każdego parametru osobno wysterowano wszystkie parametry jednocześnie. Przeprowadzono dwie ewaluacje. Wartości nastaw w ustawieniach algorytmu, w pierwszym teście (T974) pochodziły z wariantów, które osiągnęły maksimum lokalne dla wskaźnika dokładności (Tabela 6.13). Drugi test (T975) przyjął wartości ustawień z iteracji, które uzyskały największą wartość współczynnika korelacji (Tabela 6.14). Na Rysunku 6.25 zestawiono wyniki osiągniętych wartości TPR oraz FPR dla przeprowadzonych badań oraz ogólnego punktu odniesienia (T221).



Rysunek 6.25. Układ współrzędnych TPR-FPR – wyniki ewaluacji modyfikacji wszystkich parametrów jednocześnie, wykorzystując wyniki uzyskane z pojedynczych ewaluacji.

Wyniki zaprezentowane na Rysunku 6.25 wskazują na znaczącą poprawę jakości klasyfikacji narzędzia względem ogólnego punktu odniesienia. Na płaszczyźnie TPR-FPR dla obu ewaluacji współczynniki przyjęły wartości bliskie klasyfikatorowi idealnemu (TPR = 100%; FPR = 0%). Wartości liczbowe metryk przestawiono w Tabeli 6.15. Istotne jest to, że każda wartość uzyskała lepsze wartości względem punktu odniesienia. Liczba próbek zaklasyfikowanych poprawnie (TP, TN) wzrosła, zaś liczba ocen fałszywych (FP, FN) zmalała. Współczynnik korelacji wzrósł znacząco. Wynikło to z poprawy oznaczeń w klasie negatywnej (Ściek), która była grupą mniej liczną. Wartość dokładności również wzrosła, jednak poprawa nie jest tak znacząca, ponieważ obserwowany jest efekt maskowania mający swoje źródło w różnej liczności klas. Różnica pomiędzy wariantem T974 oraz T975 jest pomijalna – ten drugi oznaczył jedną próbkę mniej ze zbioru treningowego.

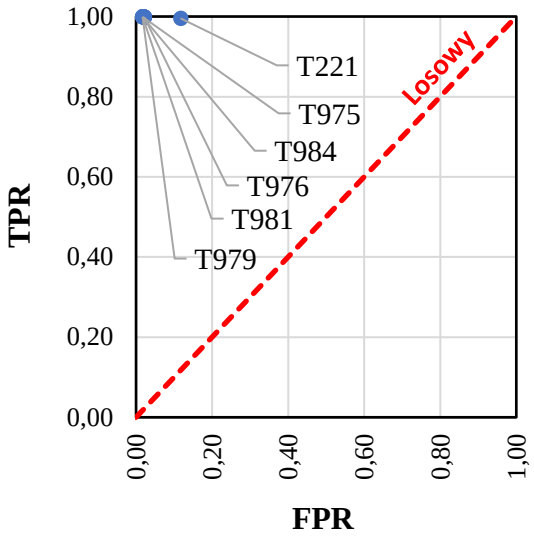
Tabela 6.15. Wyniki ewaluacji modyfikacji wszystkich parametrów jednocześnie, wykorzystując wyniki uzyskane z pojedynczych ewaluacji.

Testowany wariant	Numer testu	MCC	ACC	ASR	TP	TN	FP	FN
Ogólny punkt odniesienia	T221	0,8651	99,20%	94,57%	7090	196	26	33
Nastawy dla maksymalnej dokładności (ACC)	T974	0,9758	99,85%	95,64%	7188	229	4	7
Nastawy dla maksymalnego współczynnika korelacji Matthewsa (MCC)	T975	0,9758	99,85%	95,62%	7187	229	4	7

6.6. Ograniczenie kosztochłonnych analiz fizykochemicznych

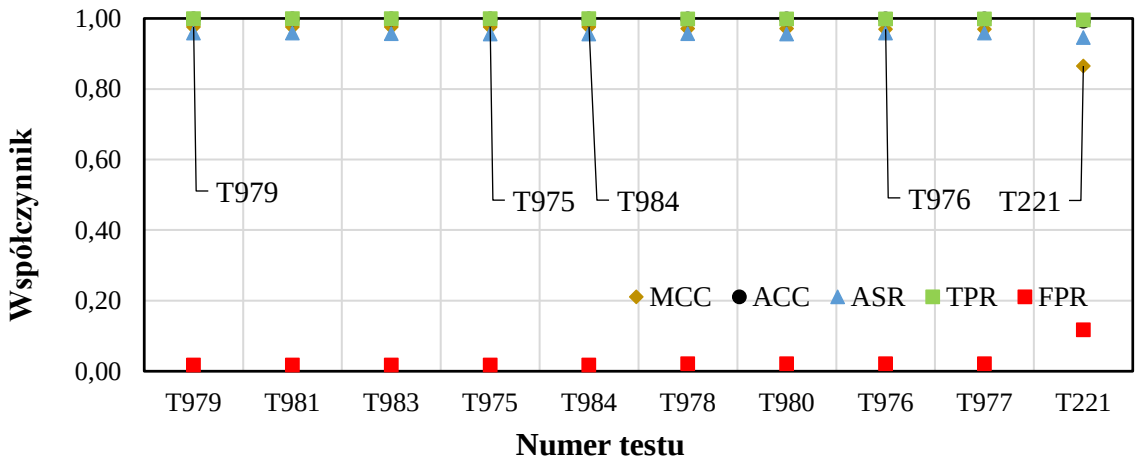
Analizy kosztochłonne to te, do których wykonania zużywane są odczynniki lub materiały jednorazowe. Badania sensoryczne polegające na porównaniu próbki ze wzorcem (kolor, wygląd, zapach) nie wymagają żadnych środków zużywalnych. Analizy pH, gęstości oraz lepkości wykonywane są z wykorzystywaniem aparatury badawczej, która poza utrzymaniem sprzętu nie generuje kosztów proporcjonalnych do liczby analiz. Oznaczanie stężeń metodami miareczkowymi (wolnego chloru, nadtlenku wodoru) wymaga używania drogich odczynników, które charakteryzują się odpowiednią czystością (np. CZDA) oraz posiadają certyfikaty jakości. Wyznaczanie suchej pozostałości wymaga stosowania jednorazowych szalek aluminiowych oraz sączków. Nie bez znaczenia pozostaje zużycie energii – rozpuszczalniki muszą być odparowane w odpowiednio krótkim czasie. Tak więc, trzy ostatnie procedury laboratoryjne (analizy miareczkowe oraz oznaczanie suchej pozostałości) zostały wskazane jako kosztochłonne.

Wariantem, który charakteryzuje się największą redukcją kosztów jest ten, który w ocenie klasyfikacji nie wymaga, żadnej ze wskazanych analiz. Postanowiono jednak przetestować każdą możliwość. Analogicznie do grup dyskretnych, stworzono matrycę kombinatoryczną (Tabela 6.16 – patrz strona 109), która uwzględnia wszystkie 8 możliwości. Wartości aktywnych parametrów wskazano na te, dla których zanotowano najwyższą wartość współczynnika korelacji, zgodnie z zestawieniem zawartym w Tabeli 6.14. Wyniki ewaluacji przedstawione zostały na Rysunku 6.26. Naniesiony został również punkt odniesienia (T221) oraz wariant z modyfikacji wszystkich parametrów jednocześnie (T975).



Rysunek 6.26. Układ współrzędnych TPR-FPR – wyniki ewaluacji modyfikacji użycia parametrów analiz kosztochłonnych.

Zauważane jest silne zgrupowanie punktów testowych (Rysunek 6.26), tak więc wpływ użycia parametrów analiz kosztochłonnych na jakość klasyfikacji był znikomy. Najmniejszą wartość dokładności zanotowano, gdy jedynie analiza wolnego chloru była aktywna (T977, ACC_{MIN} = 99,81%), zaś największą dla wariantu uwzględniającego tylko parametr suchej pozostałości (T979, ACC_{MAX} = 99,85%). Dla każdej wersji testu zaobserwowano wysokie wartości MCC, ACC, ASR, TPR (Rysunek 6.27). Istotna jest wysoka wartość oraz mała zmienność współczynnika korelacji (MCC_{MIN} = 0,9693; MCC_{MAX} = 0,9758). Analogicznie wskaźnik oznaczeń fałszywie pozytywnych pozostał na niskim poziomie (FPR_{MIN} = 1,72%; FPR_{MAX} = 2,15%). Należy zaznaczyć, że warianty, które nie uwzględniały parametru analizy suchej pozostałości uzyskały wartości FPR wyższe (gorsze) niż test referencyjny (T975).



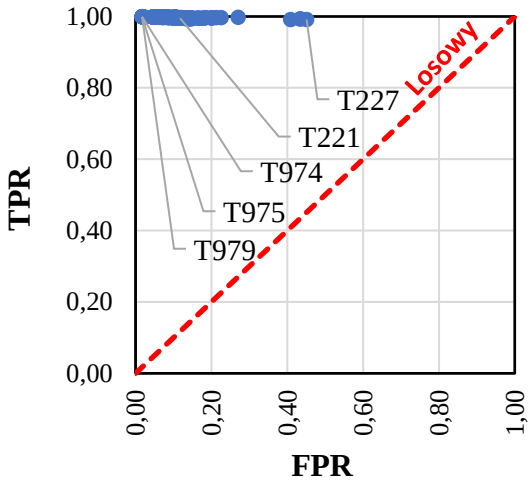
Rysunek 6.27. Profile wskaźników oceny klasyfikacji zarejestrowane podczas modyfikacji użycia parametrów analiz kosztochłonnych.

Tabela 6.16. Kombinatoryczna matryca testów użycia parametrów odpowiadających analizom kosztochłonnych, wyniki ich ewaluacji oraz warianty porównawcze T975 i T221.

Numer testu	Wolny chlor	Nadtlenek	Sucha pozostałość	MCC	ACC	ASR	TP	TN	FP	FN
T979	Nie	Nie	Tak	0,9758	99,85%	95,84%	7204	229	4	7
T981	Tak	Nie	Tak	0,9758	99,85%	95,82%	7202	229	4	7
T983	Nie	Tak	Tak	0,9758	99,85%	95,65%	7189	229	4	7
T984	Tak	Tak	Tak	0,9758	99,85%	95,62%	7187	229	4	7
T978	Nie	Tak	Nie	0,9714	99,83%	95,65%	7188	228	5	8
T980	Tak	Tak	Nie	0,9714	99,82%	95,62%	7186	228	5	8
T976	Nie	Nie	Nie	0,9693	99,81%	95,84%	7202	228	5	9
T977	Tak	Nie	Nie	0,9693	99,81%	95,82%	7200	228	5	9
T975	Tak	Tak	Tak	0,9758	99,85%	95,62%	7187	229	4	7
T221	Tak	Tak	Tak	0,8651	99,20%	94,57%	7090	196	26	33

6.7. Podsumowanie trenowania algorytmu

Na Rysunku 6.28 naniesiono wszystkie punkty testowe jakie otrzymano w trakcie badań optymalizacyjnych parametrów sterujących algorytmem.



Rysunek 6.28. Układ współrzędnych TPR-FPR – wyniki wszystkich przeprowadzonych ewaluacji algorytmu klasyfikacji.

Zauważalne jest to, że modyfikacje parametrów głównie dawała odpowiedź w postaci zmiany wartości wskaźnika próbek fałszywie pozytywnych (FPR). Wpływ modyfikacji nastaw obserwowano w obu kierunkach względem punktu odniesienia. Uzyskano wyniki znacznie lepsze (T979, Tabela 6.17) oraz znacznie gorsze (T227, Tabela 6.3).

W Tabeli 6.17 zestawione zostały warianty testów, które stanowią wynik badań ewaluacji nastaw algorytmu (trenowania modelu uczenia maszynowego). Wytypowane zostały trzy testy, które uzyskały najlepsze wartości wskaźników oceny jakości klasyfikacji, czwarty test to ogólny punkt odniesienia. W każdym przypadku optymalnego wysterowania obserwowane były lepsze wartości współczynników względem punktu odniesienia. Istotnym jest fakt, że suma oznaczeń fałszywych (FP + FN) była taka sama w każdym przypadku. Liczba próbek prawdziwie negatywnych (TN) również nie ulegała zmianie. Natomiast zanotowano zmiany w liczności oznaczeń prawdziwie pozytywnych – maksimum zostało osiągnięte dla test T979.

Warianty testów, które były jednoczesnym wysterowaniem wszystkich parametrów algorytmu (Tabela 6.15) oraz te iteracje, które badano w ramach redukcji analiz kosztochłonnych (Tabela 6.16) uzyskały lepsze wartości metryk oceny klasyfikacji niż obserwowano w pojedynczych modyfikacjach parametrów (Tabela 6.13).

Tabela 6.17. Podsumowanie ewaluacji modyfikacji parametrów algorytmu.

Testowany parametr	Numer testu	MCC	ACC	ASR	TP	TN	FP	FN
Ogólny punkt odniesienia	T221	0,8651	99,20%	94,57%	7090	196	26	33
Nastawy dla maksymalnej dokładności (ACC)	T974	0,9758	99,85%	95,64%	7188	229	4	7
Nastawy dla maksymalnego współczynnika korelacji Matthewsa (MCC)	T975	0,9758	99,85%	95,62%	7187	229	4	7
Nastawy optymalne bez analiz kosztochłonnych	T979	0,9758	99,85%	95,84%	7204	229	4	7

7. Dyskusja wyników

W pracy doktorskiej podjęte zostały trzy zagadnienia dotyczące stosowania narzędzi statystycznych w analizach jakościowych (obszary produkcyjne) w przemyśle chemicznym. Dotyczyły one możliwości wykorzystania w badaniach dostępnego sprzętu oraz oprogramowania biurowego, oceny czy zbudowany algorytm klasyfikacyjny uzyska dokładność lepszą niż losowa oraz czy istnieje potencjał wdrożeniowy opracowanego narzędzia. Poniższą część rozprawy poświęcono dyskusji otrzymanych wyników.

Limity liczby próbek uwzględnianych w obliczeniach algorytmu klasyfikacji

Parametr ograniczający liczbę próbek, która była pobierana w celu dokonania klasyfikacji posiada charakterystykę (Rysunek 6.2), która wskazuje, że im większa liczebność populacji, tym dokładność oceny jest lepsza. W omawianych wynikach ujawniało się to poprzez malejący współczynnik oznaczeń fałszywie pozytywnych (wskaźnik FPR). Osiągany był poziom minimum, po którym zwiększenie limitu nie owocowała lepszą dokładnością. Zauważalny jest negatywny wpływ limitu minimalnych liczebności próbek na jakość klasyfikacji (Rysunek 6.3, Rysunek 6.4, Rysunek 6.5). Model testowania uwzględniał stałość maksymalnej ogólnej liczby próbek. Tak więc w przypadku zwiększania dolnego limitu, liczba próbek uwzględniona w klasyfikacji malała. Matryca testów powinna zawierać warianty modyfikujące minimalną i maksymalną liczbę próbek jednocześnie. W przypadku sterowania parametrami oddzielnie dla każdej z klas zaburzana była pierwotna proporcja pomiędzy nimi. Ponadto ze względu na małą liczność klasy negatywnej, nawet niska wartość limitu powodowała drastyczny spadek w możliwości przypisania klasy do analizowanej próbki (wskaźnik ASR, Rysunek 6.5). Tak więc jedynie parametr ograniczający ogólną liczbę pozwalał na sterowanie jakością algorytmu (Rysunek 6.1).

Dyskretne parametry algorytmu

W kombinatorycznym testowaniu parametrów dyskretnych algorytmu (Rysunek 6.8) zauważalny jest głównie negatywny wpływ zastosowania korekcji braku wystąpień obserwacji (tzw. wygładzenie Laplace'a). W opracowanej aplikacji nie ma możliwości sterowania zmiennymi w procesie korekcji – jedynie użycie jej bądź nie. Tak więc w analizowanym wdrożeniu najprostsza metoda wygładzenia Laplace'a (inkrementacja licznika oraz mianownika o jeden) nie przyniosła pożądanych efektów zwiększenia dokładności klasyfikacji. Uzasadnione wydaje się być również postulowanie, że prawdopodobieństwa cząstkowe są istotne na tyle, że próbna korekcji ich przynosi efekt negatywny. Niemniej jednak, zaobserwowano pozytywny wpływ wyłączenia ze zbioru treningowego próbek, które zostały ocenione warunkowo jako pozytywne (zawierały zalecenia). Zmiana sposobu dyskretyzacji na ten oparty o odchylenie standardowe nie przyniósł widocznych zmian w wartościach metryk oceniających algorytm. Tak więc sposób podziału zakresu wartości rzeczywistych na równe przedziały okazał się satysfakcjonujący, a co istotne wymaga on mniej operacji matematycznych.

Analizy pH, gęstości oraz lepkości

Trzy analizy fizykochemiczne, które zostały uznane za nie generujące kosztów – obciążenie budżetu laboratorium nie zależy od liczby zanalizowanych próbek. W kwestii optymalizacji algorytmu szczególnie zauważalny jest wpływ dyskretyzacji wyników pH na jakość oceny klasy negatywnej (Rysunek 6.9). Optymalna nastawa obniżyła wartość współczynnika liczby próbek fałszywie pozytywnych o ponad 7 punktów procentowych względem punktu odniesienia. Parametry odpowiadające analizom gęstości (Rysunek 6.11) oraz lepkości (Rysunek 6.12) również wpływały na polepszenie tego wskaźnika, lecz już nie w tak znacznym stopniu, odpowiednio o 1,35 pp. oraz o 0,4 pp. Sugeruje to, że pH jest parametrem kluczowym, który wskazuje na przydatność partii do dalszej produkcji. O ile istotności pozostałych dwóch analiz, na bazie przeprowadzonych testów, nie można wykluczyć, o tyle wpływ zmiany dyskretyzacji ich wartości jest mniej zauważalny. W przypadku pH oraz gęstości obserwowano wysokie wartości nastaw w wariantach optymalnych (powyżej 500). Ważny jest również fakt, że w grupie tych parametrów, w żadnym wariancie dyskretyzacji wyników oznaczenia lepkości, wartość współczynnika korelacji nie wzrosła znacząco względem punktu odniesienia (Tabela 6.13, Tabela 6.14).

Analizy stężenia wolnego chloru, stężenia nadtlenku wodoru, suchej pozostałości

Każda ze wskazanych analiz jest specyficzna dla jednego z trzech typów produktów. Wewnętrzne specyfikacje określają te badania jako krytyczne dla dopuszczenia produktu do sprzedaży. Ponadto, na potrzeby niniejszych badań, zostały one wytypowane jako kosztochłonne – redukcja liczby analiz zmniejszy koszty funkcjonowania laboratorium fizykochemicznego. Cechą wspólną dla optymalizacji tych parametrów była mała zależność pomiędzy liczbą poziomów dyskretnych, a jakością klasyfikacji. Zauważalna różnica pomiędzy punktem odniesienia, a optymalną nastawą obserwowana była jedynie w liczbie próbek prawdziwie pozytywnych (Tabela 6.12, Tabela 6.13) – liczność klasy sprawiła, że zmiana współczynnika dokładności nie przekroczyła 0,3 pp. Zmiany współczynnika korelacji również nie wskazały na silną korelację z jakością przypisywania klasy, jak to miało miejsce w przypadku analizy pH.

Interesującym jest fakt, że niezastosowanie w algorytmie klasyfikacji żadnych wyników z analiz krytycznych nie pogorszyło jakości klasyfikacji (Tabela 6.16). Na bazie tych obserwacji, możliwe jest wnioskowanie, że pozostałe parametry są wystarczające do dokonania klasyfikacji. W fazie koncepcyjnej estymowana była wysoka istotność tychże parametrów, ponieważ stanowiły one podstawę do dopuszczenia wyrobu do sprzedaży. Stabilność produktów na bazie podchlorynu sodu oraz nadtlenku wodoru zależy silnie od pH. Tak więc w opracowanym modelu statystycznym, najprawdopodobniej zmiany wartości tej analizy są wrażliwe na tyle, że możliwe jest pominięcie paramentów wyrażających stężenia. Natomiast, możliwe jest, że analiza suchej pozostałości koreluje z analizą gęstości. Głównymi składnikami pozostającymi po odparowaniu rozpuszczalników są surfaktanty oraz chlorek sodu. Optymalny wariant testów kombinatorycznych przeprowadzonych w celu ograniczenia kosztów badań, zawierał jedynie analizę suchej pozostałości. Jednak, zmiany w liczbie próbek pomiędzy poszczególnymi wariantami są pomijalne. W przypadku próbek oznaczonych prawdziwie (TP, TN) jest to różnica mniejsza od 15, a w przypadku próbek oznaczonych fałszywie (FP, FN) różnica wynosi 1.

Parametry dyskretne – cechy partii produktu

Najistotniejszym parametrem dyskretnym w kategorii cech partii produktu okazała się informacja o instalacji produkcyjnej (Rysunek 6.21). Park maszynowy zakładu produkcyjnego pozwala na wytwarzanie tego samego półproduktu na różnych instalacjach. Różnią się one znacząco poziomem automatyzacji (np. dozowanie z wykorzystaniem wagi o dokładności ±2 kg, dozowanie z wykorzystaniem przepływomierzy o dokładności ±150 g), wiekiem oraz rodzajem armatury. Tak więc, miejsce wytwarzania może mieć wpływ na dozowanie surowców oraz procesy pomocnicze (np. homogenizacja), a tym samym na parametry jakościowe wyrobu.

Parametry dyskretne – porównanie produktu ze wzorcem

W analizach sensorycznych polegających na porównaniu próbki ze wzorcem najmniej istotną okazała się analiza zapachu (Rysunek 6.23). Najlepsze wyniki zaś uzyskano dla wariantów, w których jednocześnie aktywne były parametry wyglądu i koloru. Aspekt względnie niskiej istotności oceny zapachy należy tłumaczyć bogatym portfolio stosowanych kompozycji perfumujących, a tym samym szerokiego spektrum możliwości korekcji i zawrotów do produktów o tej samej bazie surowcowej. Silny pozytywny wpływ połączenia analizy wyglądu (np. zanieczyszczenia drobinami stałymi, zmętnienie) oraz koloru (ocena jedynie barwy), wynika z faktu, że niezgodność tylko jednego z tych parametrów może zostać skorygowana. Natomiast rozbieżność w obu parametrach oznacza wystąpienie więcej niż jednego problemu w trakcie produkcji. Niemniej jednak, wyłączenie wszystkich trzech analiz z obliczeń, skutkuje wzrostem liczby próbek fałszywie pozytywnych aż o 15 punktów procentowych.

Działania optymalizacyjne w obszarze inżynierii chemicznej

Przedstawiona praca badawcza udowodniła, że działania w obszarach laboratoriów chemicznych mogą być optymalizowane z wykorzystaniem metod interdyscyplinarnych. Analiza potrzeb procesu produkcyjnego (Rozdział 5.2.1 „Dobór elementu procesu kontroli jakości") oraz dostępnych narzędzi (Rozdział 4.3 „Metody klasyfikacji binarnej") pozwoliła na opracowanie nowego rozwiązania, które może zastąpić personel laboratorium w podejmowaniu decyzji – zgodność pomiędzy odpowiedzią algorytmu, a decyzją inżyniera chemika wynosiła nawet 99,85% (Tabela 6.17). Wykazano, że przedmiot badań optymalizacji działań kontroli jakości fizykochemicznej nie skupia się wyłącznie na procesach chemicznych, a na szerokiej interdyscyplinarnej analizie potrzeb oraz dostępnych technologii. Opracowane narzędzie pozwala na optymalizację kosztów oraz zwiększenie bezpieczeństwa załogi produkcyjnej.

8. Wnioski

Wyniki badań przeprowadzonych w ramach przedstawionej pracy doktorskiej doprowadziły do licznych wniosków, które przedstawiono w niniejszym rozdziale.

1) Udowodniony został potencjał wdrożeniowy opracowanego algorytmu. Dokładność przypisania klasy przekroczyła wymagany przez zakład produkcyjny poziom 95%, przy zachowaniu kryterium przyznania ocen minimum 80% próbek ze zbioru treningowego.
2) Potwierdzone została możliwościami wykorzystania narzędzi statystycznych w analizach jakościowych (cel badawczy pracy doktorskiej), w obszarach produkcyjnych. Dokładność opracowanego narzędzia klasyfikującego była większa od 50% – a tym samym lepsza od losowego przypisania klasy. W populacji wszystkich przeprowadzonych testów osiągnięto najmniejszą dokładność na poziomie 97,85%.
3) Możliwe jest opracowanie narzędzia wykorzystującego model uczenia maszynowego, pracującego na komputerze osobistym, stworzonego przy użyciu oprogramowania biurowego Microsoft Office w celach klasyfikacji próbek fizykochemicznych. Zatem do początkowych prac z algorytmami uczenia maszynowego nie są konieczne inwestycje w nowe oprogramowani ani w sprzęt komputerowy.
4) Wykazano, że możliwości standardowego oprogramowania biurowego są wystraczające do zaprojektowania oraz zbudowania narzędzia, które może zostać wdrożone zgodnie z obowiązującymi normami międzynarodowymi oraz branżowymi. Wykorzystanie modelu analizy ryzyka (FEMA) oraz cyklu ciągłego doskonalenie (PDCA) usprawnia budowanie koncepcji oraz określanie szczegółowych wymagań dotyczących przedmiotu wdrożenia.
5) Przeprowadzone badania potwierdziły, że algorytm naiwnego klasyfikatora Bayesa zapewnia wystarczającą jakość oceny w produkcyjnej kontroli jakości.
6) Prace optymalizacyjne wykazały, że liczba badań fizykochemicznych może zostać zredukowana. Tak więc, narzędzie statystyczne pozwala na zmniejszenie ilości pracy inżynierów chemików przy ocenie próbek produkcyjnych oraz ograniczenie kosztów

zużycia materiałów i odczynników. Ponadto, ograniczenie kontaktu z odczynnikami może zwiększać bezpieczeństwo personelu laboratoryjnego (np. fenoloftaleina znajduje się na liście substancji rakotwórczych, zaś chloroform ma kategorię szkodliwości na rozrodczość).

7) Pomimo faktu, że liczność klas w zbiorze treningowym nie była symetryczna, to nie zaobserwowano znaczącej różnicy w końcowych nastawach algorytmu dobranych na podstawie dokładności (wskaźnika niezbalansowanego) oraz współczynnika korelacji Matthewsa (wskaźnika zbalansowanego).

9. Literatura

[1] Primi A, Toselli M. A global perspective on industry 4.0 and development: new gaps or opportunities to leapfrog?, *Journal of Economic Policy Reform*, 2020, 23, 371–89. DOI:10.1080/17487870.2020.1727322.

[2] Szpadzik D. Quo vadis? Przemysł 4.0, *Przemysł Chemiczny*, 2021, 1, 5–6. DOI:10.15199/62.2021.12.1.

[3] Mohanta B, Nanda P, Patnaik S. Management of V.U.C.A. (Volatility, Uncertainty, Complexity and Ambiguity) Using Machine Learning Techniques in Industry 4.0 Paradigm, 2020, p. 1–24. DOI:10.1007/978-3-030-25778-1_1.

[4] Jasperneite J, Sauter T, Wollschlaeger M. Why We Need Automation Models: Handling Complexity in Industry 4.0 and the Internet of Things, *IEEE Industrial Electronics Magazine*, 2020, 14, 29–40. DOI:10.1109/MIE.2019.2947119.

[5] Trattner A, Hvam L, Forza C, Herbert-Hansen ZNL. Product complexity and operational performance: A systematic literature review, *CIRP J Manuf Sci Technol*, 2019, 25, 69–83. DOI:10.1016/j.cirpj.2019.02.001.

[6] Mocker M, Ross J. The Problem with Product Proliferation, *Harv Bus Rev*, 2017, 95, 104–10.

[7] Mourtzis D. The mass personalization of global networks. *Design and Operation of Production Networks for Mass Personalization in the Era of Cloud Technology*, Elsevier, 2022, p. 79–116. DOI:10.1016/B978-0-12-823657-4.00006-3.

[8] Vidal GH, Hernández JRC. Study of the effects of complexity on the manufacturing sector, *Production Engineering*, 2021, 15, 69–78. DOI:10.1007/s11740-020-01014-2.

[9] Romero D, Gaiardelli P, Powell D, Wuest T, Thürer M. Total Quality Management and Quality Circles in the Digital Lean Manufacturing World, 2019, p. 3–11. DOI:10.1007/978-3-030-30000-5_1.

[10] ISO 9001:2015 Quality management systems — Requirements. ISO 9001:2015 Quality management systems — Requirements 2015.

[11] Rozporządzenie Parlamentu Europejskiego i Rady (UE) nr 528/2012 z dnia 22 maja 2012 r. w sprawie udostępniania na rynku i stosowania produktów biobójczych Tekst mający znaczenie dla EOG, Parlament Europejski, Rada Unii Europejskiej, 2012.

[12] Barkman W. In-process quality control for manufacturing, CRC Press, 1989.

[13] Ribeiro J, Lima R, Eckhardt T, Paiva S. Robotic Process Automation and Artificial Intelligence in Industry 4.0 – A Literature review, *Procedia Comput Sci*, 2021, 181, 51–8. DOI:10.1016/j.procs.2021.01.104.

[14] Guo D, Ling S, Li H, Ao D, Zhang T, Rong Y, et al. A framework for personalized production based on digital twin, blockchain and additive manufacturing in the context of Industry 4.0. *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, IEEE, 2020, p. 1181–6. DOI:10.1109/CASE48305.2020.9216732.

[15] Montgomery DC. Introduction to statistical quality control, John Wiley & Sons, 2020.

[16] Escobar CA, Morales-Menendez R. Machine learning techniques for quality control in high conformance manufacturing environment, *Advances in Mechanical Engineering*, 2018, 10, 168781401875551. DOI:10.1177/1687814018755519.

[17] Pandey D, Kulkarni MS, Vrat P. A methodology for joint optimization for maintenance planning, process quality and production scheduling, *Comput Ind Eng*, 2011, 61, 1098–106. DOI:10.1016/j.cie.2011.06.023.

[18] Zhou J, Gandomi AH, Chen F, Holzinger A. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics, *Electronics (Basel)*, 2021, 10, 593. DOI:10.3390/electronics10050593.

[19] Baum ZJ, Yu X, Ayala PY, Zhao Y, Watkins SP, Zhou Q. Artificial Intelligence in Chemistry: Current Trends and Future Directions, *J Chem Inf Model*, 2021, 61, 3197–212. DOI:10.1021/acs.jcim.1c00619.

[20] Keith JA, Vassilev-Galindo V, Cheng B, Chmiela S, Gastegger M, Müller K-R, et al. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems, *Chem Rev*, 2021, 121, 9816–72. DOI:10.1021/acs.chemrev.1c00107.

[21] Godfrey AG, Michael SG, Sittampalam GS, Zahoránszky-Köhalmi G. A Perspective on Innovating the Chemistry Lab Bench, *Front Robot AI*, 2020, 7. DOI:10.3389/frobt.2020.00024.

[22] Westermayr J, Gastegger M, Schütt KT, Maurer RJ. Perspective on integrating machine learning into computational chemistry and materials science, *J Chem Phys*, 2021, 154. DOI:10.1063/5.0047760.

[23] Villarrel LP. Are Viruses Alive?, *Sci Am*, 2004, 291, 100–5.

[24] What is machine learning?, *IBM*, n.d. https://www.ibm.com/topics/machine-learning (dostęp 03/23/2023).

[25] Brown S. Machine learning, explained, *MIT Sloan School of Management*, 2021. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained (dostęp 03/29/2023).

[26] Spector L. Evolution of artificial intelligence, *Artif Intell*, 2006, 170, 1251–3. DOI:10.1016/j.artint.2006.10.009.

[27] Bini SA. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care?, *J Arthroplasty*, 2018, 33, 2358–61. DOI:10.1016/j.arth.2018.02.067.

[28] Chassagnon G, Vakalopolou M, Paragios N, Revel M-P. Deep learning: definition and perspectives for thoracic imaging, *Eur Radiol*, 2020, 30, 2021–30. DOI:10.1007/s00330-019-06564-3.

[29] Mahesh B. Machine Learning Algorithms - A Review, *International Journal of Science and Research*, 2020, 9, 381–6.

[30] LeCun Y, Bengio Y, Hinton G. Deep learning, *Nature*, 2015, 521, 436–44. DOI:10.1038/nature14539.

[31] Popper J, Hermann J, Cui K, Bergweiler S, Weyer S, Ruskowski M, et al. Artificial intelligence across industries - IEC Whitepaper, 2018.

[32] Samuel AL. Some Studies in Machine Learning Using the Game of Checkers, *IBM J Res Dev*, 1959, 3, 210–29. DOI:10.1147/rd.33.0210.

[33] El Naqa I, Murphy MJ. What Is Machine Learning? *Machine Learning in Radiation Oncology*, Cham, Springer International Publishing, 2015, p. 3–11. DOI:10.1007/978-3-319-18305-3_1.

[34] Helm JM, Swiergosz AM, Haeberle HS, Karnuta JM, Schaffer JL, Krebs VE, et al. Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions, *Curr Rev Musculoskelet Med*, 2020, 13, 69–76. DOI:10.1007/s12178-020-09600-8.

[35] Burkart N, Huber MF. A Survey on the Explainability of Supervised Machine Learning, *Journal of Artificial Intelligence Research*, 2021, 70, 245–317. DOI:10.1613/jair.1.12228.

[36] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning, New York, NY, Springer New York, 2009. DOI:10.1007/978-0-387-84858-7.

[37] Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size, *PLoS One*, 2019, 14, e0224365. DOI:10.1371/journal.pone.0224365.

[38] Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners, *IEEE Trans Pattern Anal Mach Intell*, 1991, 13, 252–64. DOI:10.1109/34.75512.

[39] Kadhim AI. Survey on supervised machine learning techniques for automatic text classification, *Artif Intell Rev*, 2019, 52, 273–92. DOI:10.1007/s10462-018-09677-1.

[40] Ali MM, Paul BK, Ahmed K, Bui FM, Quinn JMW, Moni MA. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison, *Comput Biol Med*, 2021, 136, 104672. DOI:10.1016/j.compbiomed.2021.104672.

[41] Jan M, Awan AA, Khalid MS, Nisar S. Ensemble approach for developing a smart heart disease prediction system using classification algorithms, *Research Reports in Clinical Cardiology*, 2018, Volume 9, 33–45. DOI:10.2147/RRCC.S172035.

[42] Althnian A, AlSaeed D, Al-Baity H, Samha A, Dris A Bin, Alzakari N, et al. Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain, *Applied Sciences*, 2021, 11, 796. DOI:10.3390/app11020796.

[43] Radja M, Emanuel AWR. Performance Evaluation of Supervised Machine Learning Algorithms Using Different Data Set Sizes for Diabetes Prediction. *2019 5th International Conference on Science in Information Technology (ICSITech)*, IEEE, 2019, p. 252–8. DOI:10.1109/ICSITech46713.2019.8987479.

[44] Matloff N. Statistical Regression and Classification: From Linear Models to Machine Learning, CRC Press, 2017.

[45] Coenen L, Verbeke W, Guns T. Machine learning methods for short-term probability of default: A comparison of classification, regression and ranking methods, *Journal of the Operational Research Society*, 2022, 73, 191–206. DOI:10.1080/01605682.2020.1865847.

[46] Ray S. A Quick Review of Machine Learning Algorithms. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, IEEE, 2019, p. 35–9. DOI:10.1109/COMITCon.2019.8862451.

[47] Maulud D, Abdulazeez AM. A Review on Linear Regression Comprehensive in Machine Learning, *Journal of Applied Science and Technology Trends*, 2020, 1, 140–7. DOI:10.38094/jastt1457.

[48] Huang J-C, Ko K-M, Shu M-H, Hsu B-M. Application and comparison of several machine learning algorithms and their integration models in regression problems, *Neural Comput Appl*, 2020, 32, 5461–9. DOI:10.1007/s00521-019-04644-5.

[49] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, *CoRR*, 2016, abs/1602.04938.

[50] Charbuty B, Abdulazeez A. Classification Based on Decision Tree Algorithm for Machine Learning, *Journal of Applied Science and Technology Trends*, 2021, 2, 20–8. DOI:10.38094/jastt20165.

[51] Aized Amin Soofi, Arshad Awan. Classification Techniques in Machine Learning: Applications and Issues, *Journal of Basic & Applied Sciences*, 2017, 13, 459–65. DOI:10.6000/1927-5129.2017.13.76.

[52] Brown FK. Chemoinformatics: What is it and How does it Impact Drug Discovery., 1998, p. 375–84. DOI:10.1016/S0065-7743(08)61100-8.

[53] Smith EG. Machine Searching for Chemical Structures, *Science (1979)*, 1960, 131, 142–6. DOI:10.1126/science.131.3394.142.

[54] Jurs PC, Kowalski BR, Isenhour TL. Computerized learning machines applied to chemical problems. Molecular formula determination from low resolution mass spectrometry, *Anal Chem*, 1969, 41, 21–7. DOI:10.1021/ac60270a002.

[55] Wang Z, Zhang W, Liu B. Computational Analysis of Synthetic Planning: Past and Future, *Chin J Chem*, 2021, 39, 3127–43. DOI:10.1002/cjoc.202100273.

[56] Gray NAB. Artificial intelligence in chemistry, *Anal Chim Acta*, 1988, 210, 9–32. DOI:10.1016/S0003-2670(00)83874-X.

[57] Diwan P. Artificial Intelligence and the Chemical World: Expert System Applications in Chemical Analysis, Chemical Synthesis and Chemical Engineering, *IETE J Res*, 1988, 34, 223–30. DOI:10.1080/03772063.1988.11436733.

[58] Hippe Z. Problems in the application of artificial intelligence in analytical chemistry, *Anal Chim Acta*, 1983, 150, 11–21. DOI:10.1016/S0003-2670(00)85455-0.

[59] Rosenkranz HS, Mitchell CS, Klopman G. Artificial intelligence and Bayesian decision theory in the prediction of chemical carcinogens, *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 1985, 150, 1–11. DOI:10.1016/0027-5107(85)90095-8.

[60] Stephanopoulos G. Artificial intelligence in process engineering—current state and future trends, *Comput Chem Eng*, 1990, 14, 1259–70. DOI:10.1016/0098-1354(90)80006-W.

[61] Chen WL. Chemoinformatics: Past, Present, and Future, *J Chem Inf Model*, 2006, 46, 2230–55. DOI:10.1021/ci060016u.

[62] Warr WA. Some Trends in Chem(o)informatics, 2010, p. 1–37. DOI:10.1007/978-1-60761-839-3_1.

[63] Willett P. Chemoinformatics: a history, *WIREs Computational Molecular Science*, 2011, 1, 46–56. DOI:10.1002/wcms.1.

[64] López-López E, Bajorath J, Medina-Franco JL. Informatics for Chemistry, Biology, and Biomedical Sciences, *J Chem Inf Model*, 2021, 61, 26–35. DOI:10.1021/acs.jcim.0c01301.

[65] Gasteiger J. Handbook of Chemoinformatics, vol. 4, Wiley, 2003. DOI:10.1002/9783527618279.

[66] About the Journal, *ACS Publications*, 2023. https://pubs.acs.org/page/jcisd8/about.html (dostęp 04/12/2023).

[67] Journal Metrics Reports 2020, *Springer Nature*, 2023. https://www.springer.com/gp/journal-impact/chemistry (dostęp 04/12/2023).

[68] About This Journal, *Wiley Online Library*, 2023. https://onlinelibrary.wiley.com/journal/18681751 (dostęp 04/12/2023).

[69] Computation sparks chemical discovery, *Nat Commun*, 2020, 11, 4811. DOI:10.1038/s41467-020-18651-x.

[70] Artrith N, Butler KT, Coudert F-X, Han S, Isayev O, Jain A, et al. Best practices in machine learning for chemistry, *Nat Chem*, 2021, 13, 505–8. DOI:10.1038/s41557-021-00716-z.

[71] Bender A, Schneider N, Segler M, Patrick Walters W, Engkvist O, Rodrigues T. Evaluation guidelines for machine learning tools in the chemical sciences, *Nat Rev Chem*, 2022, 6, 428–42. DOI:10.1038/s41570-022-00391-9.

[72] Trinh C, Meimaroglou D, Hoppe S. Machine Learning in Chemical Product Engineering: The State of the Art and a Guide for Newcomers, *Processes*, 2021, 9, 1456. DOI:10.3390/pr9081456.

[73] Wang AY-T, Murdock RJ, Kauwe SK, Oliynyk AO, Gurlo A, Brgoch J, et al. Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices, *Chemistry of Materials*, 2020, 32, 4954–65. DOI:10.1021/acs.chemmater.0c01907.

[74] Klambauer G, Hochreiter S, Rarey M. Machine Learning in Drug Discovery, *J Chem Inf Model*, 2019, 59, 945–6. DOI:10.1021/acs.jcim.9b00136.

[75] Wei J, Chu X, Sun X, Xu K, Deng H, Chen J, et al. Machine learning in materials science, *InfoMat*, 2019, 1, 338–58. DOI:10.1002/inf2.12028.

[76] Patel L, Shukla T, Huang X, Ussery DW, Wang S. Machine Learning Methods in Drug Discovery, *Molecules*, 2020, 25, 5277. DOI:10.3390/molecules25225277.

[77] Dral PO, Barbatti M. Molecular excited states through a machine learning lens, *Nat Rev Chem*, 2021, 5, 388–405. DOI:10.1038/s41570-021-00278-1.

[78] Bartók AP, De S, Poelking C, Bernstein N, Kermode JR, Csányi G, et al. Machine learning unifies the modeling of materials and molecules, *Sci Adv*, 2017, 3. DOI:10.1126/sciadv.1701816.

[79] von Lilienfeld OA, Müller K-R, Tkatchenko A. Exploring chemical compound space with quantum-based machine learning, *Nat Rev Chem*, 2020, 4, 347–58. DOI:10.1038/s41570-020-0189-9.

[80] Zhang Y, Xu X. Machine learning lattice constants for spinel compounds, *Chem Phys Lett*, 2020, 760, 137993. DOI:10.1016/j.cplett.2020.137993.

[81] Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science, *NPJ Comput Mater*, 2019, 5, 83. DOI:10.1038/s41524-019-0221-0.

[82] Lv C, Zhou X, Zhong L, Yan C, Srinivasan M, Seh ZW, et al. Machine Learning: An Advanced Platform for Materials Development and State Prediction in Lithium-Ion Batteries, *Advanced Materials*, 2022, 34, 2101474. DOI:10.1002/adma.202101474.

[83] Moosavi SM, Jablonka KM, Smit B. The Role of Machine Learning in the Understanding and Design of Materials, *J Am Chem Soc*, 2020, 142, 20273–87. DOI:10.1021/jacs.0c09105.

[84] Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine Learning in Drug Discovery: A Review, *Artif Intell Rev*, 2022, 55, 1947–99. DOI:10.1007/s10462-021-10058-4.

[85] Reker D. Practical considerations for active machine learning in drug discovery, *Drug Discov Today Technol*, 2019, 32–33, 73–9. DOI:10.1016/j.ddtec.2020.06.001.

[86] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development, *Nat Rev Drug Discov*, 2019, 18, 463–77. DOI:10.1038/s41573-019-0024-5.

[87] Patel V, Shah M. Artificial intelligence and machine learning in drug discovery and development, *Intelligent Medicine*, 2022, 2, 134–40. DOI:10.1016/j.imed.2021.10.001.

[88] Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, Cedrón F, Novoa FJ, Carballal A, et al. A review on machine learning approaches and trends in drug discovery, *Comput Struct Biotechnol J*, 2021, 19, 4538–58. DOI:10.1016/j.csbj.2021.08.011.

[89] Chandrasekaran SN, Ceulemans H, Boyd JD, Carpenter AE. Image-based profiling for drug discovery: due for a machine-learning upgrade?, *Nat Rev Drug Discov*, 2021, 20, 145–59. DOI:10.1038/s41573-020-00117-w.

[90] Aykol M, Herring P, Anapolsky A. Machine learning for continuous innovation in battery technologies, *Nat Rev Mater*, 2020, 5, 725–7. DOI:10.1038/s41578-020-0216-y.

[91] Roman D, Saxena S, Robu V, Pecht M, Flynn D. Machine learning pipeline for battery state-of-health estimation, *Nat Mach Intell*, 2021, 3, 447–56. DOI:10.1038/s42256-021-00312-3.

[92] Chen C, Zuo Y, Ye W, Li X, Deng Z, Ong SP. A Critical Review of Machine Learning of Energy Materials, *Adv Energy Mater*, 2020, 10, 1903242. DOI:10.1002/aenm.201903242.

[93] Schweidtmann AM, Esche E, Fischer A, Kloft M, Repke J, Sager S, et al. Machine Learning in Chemical Engineering: A Perspective, *Chemie Ingenieur Technik*, 2021, 93, 2029–39. DOI:10.1002/cite.202100083.

[94] Dobbelaere MR, Plehiers PP, Van de Vijver R, Stevens C V., Van Geem KM. Machine Learning in Chemical Engineering: Strengths, Weaknesses, Opportunities, and Threats, *Engineering*, 2021, 7, 1201–11. DOI:10.1016/j.eng.2021.03.019.

[95] Shokry A, Medina-González S, Baraldi P, Zio E, Moulines E, Espuña A. A machine learning-based methodology for multi-parametric solution of chemical processes operation optimization under uncertainty, *Chemical Engineering Journal*, 2021, 425, 131632. DOI:10.1016/j.cej.2021.131632.

[96] Weichert D, Link P, Stoll A, Rüping S, Ihlenfeldt S, Wrobel S. A review of machine learning for the optimization of production processes, *The International Journal of Advanced Manufacturing Technology*, 2019, 104, 1889–902. DOI:10.1007/s00170-019-03988-5.

[97] Diez-Olivan A, Del Ser J, Galar D, Sierra B. Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0, *Information Fusion*, 2019, 50, 92–111. DOI:10.1016/j.inffus.2018.10.005.

[98] Min Q, Lu Y, Liu Z, Su C, Wang B. Machine Learning based Digital Twin Framework for Production Optimization in Petrochemical Industry, *Int J Inf Manage*, 2019, 49, 502–19. DOI:10.1016/j.ijinfomgt.2019.05.020.

[99] Coley CW, Green WH, Jensen KF. Machine Learning in Computer-Aided Synthesis Planning, *Acc Chem Res*, 2018, 51, 1281–9. DOI:10.1021/acs.accounts.8b00087.

[100] Haghighatlari M, Vishwakarma G, Altarawy D, Subramanian R, Kota BU, Sonpal A, et al. ChemML: A machine learning and informatics program package for the analysis, mining, and modeling of chemical and materials data, *WIREs Computational Molecular Science*, 2020, 10. DOI:10.1002/wcms.1458.

[101] Strieth-Kalthoff F, Sandfort F, Segler MHS, Glorius F. Machine learning the ropes: principles, applications and directions in synthetic chemistry, *Chem Soc Rev*, 2020, 49, 6154–68. DOI:10.1039/C9CS00786E.

[102] Fortunato ME, Coley CW, Barnes BC, Jensen KF. Data Augmentation and Pretraining for Template-Based Retrosynthetic Prediction in Computer-Aided Synthesis Planning, *J Chem Inf Model*, 2020, 60, 3398–407. DOI:10.1021/acs.jcim.0c00403.

[103] Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF. Prediction of Organic Reaction Outcomes Using Machine Learning, *ACS Cent Sci*, 2017, 3, 434–43. DOI:10.1021/acscentsci.7b00064.

[104] Meuwly M. Machine Learning for Chemical Reactions, *Chem Rev*, 2021, 121, 10218–39. DOI:10.1021/acs.chemrev.1c00033.

[105] Oliveira JCA, Frey J, Zhang S-Q, Xu L-C, Li X, Li S-W, et al. When machine learning meets molecular synthesis, *Trends Chem*, 2022, 4, 863–85. DOI:10.1016/j.trechm.2022.07.005.

[106] Hein JE. Machine learning made easy for optimizing chemical reactions, *Nature*, 2021, 590, 40–1. DOI:10.1038/d41586-021-00209-6.

[107] Zhou Z, Li X, Zare RN. Optimizing Chemical Reactions with Deep Reinforcement Learning, *ACS Cent Sci*, 2017, 3, 1337–44. DOI:10.1021/acscentsci.7b00492.

[108] Srinivasa Rao PC, Sravan Kumar AJ, Niyaz Q, Sidike P, Devabhaktuni VK. Binary chemical reaction optimization based feature selection techniques for machine learning classification problems, *Expert Syst Appl*, 2021, 167, 114169. DOI:10.1016/j.eswa.2020.114169.

[109] Brereton RG. Pattern recognition in chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 2015, 149, 90–6. DOI:10.1016/j.chemolab.2015.06.012.

[110] Houhou R, Bocklitz T. Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data, *Analytical Science Advances*, 2021, 2, 128–41. DOI:10.1002/ansa.202000162.

[111] Puthongkham P, Wirojsaengthong S, Suea-Ngam A. Machine learning and chemometrics for electrochemical sensors: moving forward to the future of analytical chemistry, *Analyst*, 2021, 146, 6351–64. DOI:10.1039/D1AN01148K.

[112] Kundu PK, Kundu M. Classification of tea samples using SVM as machine learning component of E-tongue. *2016 International Conference on Intelligent Control Power and Instrumentation (ICICPI)*, IEEE, 2016, p. 56–60. DOI:10.1109/ICICPI.2016.7859673.

[113] Tan J, Xu J. Applications of electronic nose (e-nose) and electronic tongue (e-tongue) in food quality-related properties determination: A review, *Artificial Intelligence in Agriculture*, 2020, 4, 104–15. DOI:10.1016/j.aiia.2020.06.003.

[114] Liu M, Wang M, Wang J, Li D. Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar, *Sens Actuators B Chem*, 2013, 177, 970–80. DOI:10.1016/j.snb.2012.11.071.

[115] Kuhn S, Egert B, Neumann S, Steinbeck C. Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction, *BMC Bioinformatics*, 2008, 9, 400. DOI:10.1186/1471-2105-9-400.

[116] Cobas C. NMR signal processing, prediction, and structure verification with machine learning techniques, *Magnetic Resonance in Chemistry*, 2020, 58, 512–9. DOI:10.1002/mrc.4989.

[117] Raljević D, Parlov Vuković J, Smrečki V, Marinić Pajc L, Novak P, Hrenar T, et al. Machine learning approach for predicting crude oil stability based on NMR spectroscopy, *Fuel*, 2021, 305, 121561. DOI:10.1016/j.fuel.2021.121561.

[118] Laanait N, Zhang Z, Schlepütz CM. Imaging nanoscale lattice variations by machine learning of x-ray diffraction microscopy data, *Nanotechnology*, 2016, 27, 374002. DOI:10.1088/0957-4484/27/37/374002.

[119] Zhao B, Greenberg JA, Wolter S. Application of machine learning to x-ray diffraction-based classification. In: Ashok A, Neifeld MA, Gehm ME, Greenberg JA, editors. *Anomaly Detection and Imaging with X-Rays (ADIX) III*, SPIE, 2018, p. 4. DOI:10.1117/12.2304683.

[120] Suzuki Y, Hino H, Hawai T, Saito K, Kotsugi M, Ono K. Symmetry prediction and knowledge discovery from X-ray diffraction patterns using an interpretable machine learning approach, *Sci Rep*, 2020, 10, 21790. DOI:10.1038/s41598-020-77474-4.

[121] Chen Z, Andrejevic N, Drucker NC, Nguyen T, Xian RP, Smidt T, et al. Machine learning on neutron and x-ray scattering and spectroscopies, *Chemical Physics Reviews*, 2021, 2, 031301. DOI:10.1063/5.0049111.

[122] Huang Y, Wang S, Guan Y, Maier A. Limited angle tomography for transmission X-ray microscopy using deep learning, *J Synchrotron Radiat*, 2020, 27, 477–85. DOI:10.1107/S160057752000017X.

[123] Kalinin S V., Ophus C, Voyles PM, Erni R, Kepaptsoglou D, Grillo V, et al. Machine learning in scanning transmission electron microscopy, *Nature Reviews Methods Primers*, 2022, 2, 11. DOI:10.1038/s43586-022-00095-w.

[124] Muto S, Shiga M. Application of machine learning techniques to electron microscopic/spectroscopic image data analysis, *Microscopy*, 2020, 69, 110–22. DOI:10.1093/jmicro/dfz036.

[125] Akers S, Kautz E, Trevino-Gavito A, Olszta M, Matthews BE, Wang L, et al. Rapid and flexible segmentation of electron microscopy data using few-shot machine learning, *NPJ Comput Mater*, 2021, 7, 187. DOI:10.1038/s41524-021-00652-z.

[126] Botifoll M, Pinto-Huguet I, Arbiol J. Machine learning in electron microscopy for advanced nanocharacterization: current developments, available tools and future outlook, *Nanoscale Horiz*, 2022, 7, 1427–77. DOI:10.1039/D2NH00377E.

[127] Lee B, Yoon S, Lee JW, Kim Y, Chang J, Yun J, et al. Statistical Characterization of the Morphologies of Nanoparticles through Machine Learning Based Electron Microscopy Image Analysis, *ACS Nano*, 2020, 14, 17125–33. DOI:10.1021/acsnano.0c06809.

[128] Coskun O. Separation Tecniques: CHROMATOGRAPHY, *North Clin Istanb*, 2016. DOI:10.14744/nci.2016.32757.

[129] Haddad PR, Taraji M, Szücs R. Prediction of Analyte Retention Time in Liquid Chromatography, *Anal Chem*, 2021, 93, 228–56. DOI:10.1021/acs.analchem.0c04190.

[130] Bouwmeester R, Martens L, Degroeve S. Comprehensive and Empirical Evaluation of Machine Learning Algorithms for Small Molecule LC Retention Time Prediction, *Anal Chem*, 2019, 91, 3694–703. DOI:10.1021/acs.analchem.8b05820.

[131] Ma C, Ren Y, Yang J, Ren Z, Yang H, Liu S. Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning, *Anal Chem*, 2018, 90, 10881–8. DOI:10.1021/acs.analchem.8b02386.

[132] Lebanov L, Tedone L, Ghiasvand A, Paull B. Random Forests machine learning applied to gas chromatography – Mass spectrometry derived average mass spectrum data sets for

classification and characterisation of essential oils, *Talanta*, 2020, 208, 120471. DOI:10.1016/j.talanta.2019.120471.

[133] Jain T, Boland T, Lilov A, Burnina I, Brown M, Xu Y, et al. Prediction of delayed retention of antibodies in hydrophobic interaction chromatography from sequence using machine learning, *Bioinformatics*, 2017, 33, 3758–66. DOI:10.1093/bioinformatics/btx519.

[134] Tiwari A, Bansode V, Rathore AS. Application of advanced machine learning algorithms for anomaly detection and quantitative prediction in protein A chromatography, *J Chromatogr A*, 2022, 1682, 463486. DOI:10.1016/j.chroma.2022.463486.

[135] Jirayupat C, Nagashima K, Hosomi T, Takahashi T, Tanaka W, Samransuksamer B, et al. Image Processing and Machine Learning for Automated Identification of Chemo-/Biomarkers in Chromatography–Mass Spectrometry, *Anal Chem*, 2021, 93, 14708–15. DOI:10.1021/acs.analchem.1c03163.

[136] Randazzo GM, Bileck A, Danani A, Vogt B, Groessl M. Steroid identification via deep learning retention time predictions and two-dimensional gas chromatography-high resolution mass spectrometry, *J Chromatogr A*, 2020, 1612, 460661. DOI:10.1016/j.chroma.2019.460661.

[137] Tian H, Wu D, Chen B, Yuan H, Yu H, Lou X, et al. Rapid identification and quantification of vegetable oil adulteration in raw milk using a flash gas chromatography electronic nose combined with machine learning, *Food Control*, 2023, 150, 109758. DOI:10.1016/j.foodcont.2023.109758.

[138] Harrington P. Machine Learning in Action, Manning, 2012.

[139] Institute for Statistics and Mathematics of Vienna University of Economics and Business. R-Project Documentation n.d.

[140] Institute for Statistics and Mathematics of Vienna University of Economics and Business. Machine Learning Pipelines for R n.d. https://search.r-project.org/CRAN/refmans/pipeliner/html/00Index.html (dostęp 04/25/2023).

[141] The MathWorks Inc. MATLAB for Machine Learning n.d. https://www.mathworks.com/solutions/machine-learning.html (dostęp 04/25/2023).

[142] Why TensorFlow n.d. https://www.tensorflow.org/about?hl=en (dostęp 04/25/2023).

[143] E W. Machine Learning and Computational Mathematics 2020. DOI:10.4208/cicp.OA-2020-0185.

[144] E W. The Dawning of a New Era in Applied Mathematics, *Notices of the American Mathematical Society*, 2021, 68, 1. DOI:10.1090/noti2259.

[145] Kumari R, Kr. S. Machine Learning: A Review on Binary Classification, *Int J Comput Appl*, 2017, 160, 11–5. DOI:10.5120/ijca2017913083.

[146] Bahel V, Pillai S, Malhotra M. A Comparative Study on Various Binary Classification Algorithms and their Improved Variant for Optimal Performance. *2020 IEEE Region 10 Symposium (TENSYMP)*, IEEE, 2020, p. 495–8. DOI:10.1109/TENSYMP50017.2020.9230877.

[147] DeMaris A. A Tutorial in Logistic Regression, *J Marriage Fam*, 1995, 57, 956. DOI:10.2307/353415.

[148] Zou X, Hu Y, Tian Z, Shen K. Logistic Regression Model Optimization and Case Analysis. *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, IEEE, 2019, p. 135–9. DOI:10.1109/ICCSNT47585.2019.8962457.

[149] Bartłomowicz T. Klasyfikacja nieruchomości metodą k-najbliższych sąsiadów, *Prace Naukowe Uniwersytetu Ekonomicznego We Wrocławiu*, 2010.

[150] Lopez-Bernal D, Balderas D, Ponce P, Molina A. Education 4.0: Teaching the Basics of KNN, LDA and Simple Perceptron Algorithms for Binary Classification Problems, *Future Internet*, 2021, 13, 193. DOI:10.3390/fi13080193.

[151] Loh W. Classification and regression trees, *WIREs Data Mining and Knowledge Discovery*, 2011, 1, 14–23. DOI:10.1002/widm.8.

[152] Afanador NL, Smolinska A, Tran TN, Blanchet L. Unsupervised random forest: a tutorial with case studies, *J Chemom*, 2016, 30, 232–41. DOI:10.1002/cem.2790.

[153] Shaik AB, Srinivasan S. A Brief Survey on Random Forest Ensembles in Classification Model, 2019, p. 253–60. DOI:10.1007/978-981-13-2354-6_27.

[154] Ranganathan S, Nakai K, Schonbach C. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, vol. 1, Elsevier Science, 2018.

[155] K H, Tayal S, George PM, Singla P, Kose U. Bayesian Reasoning and Gaussian Processes for Machine Learning Applications, CRC Press, 2022.

[156] Taheri S, Mammadov M. Learning the naive Bayes classifier with optimization models, *International Journal of Applied Mathematics and Computer Science*, 2013, 23, 787–95. DOI:10.2478/amcs-2013-0059.

[157] Yang Y, Webb GI. On Why Discretization Works for Naive-Bayes Classifiers, 2003, p. 440–52. DOI:10.1007/978-3-540-24581-0_37.

[158] Abraham R, Simha JB, Iyengar SS. A comparative analysis of discretization methods for Medical Datamining with Naive Bayesian classifier. *9th International Conference on Information Technology (ICIT'06)*, IEEE, 2006, p. 235–6. DOI:10.1109/ICIT.2006.5.

[159] Dougherty J, Kohavi R, Sahami M. Supervised and Unsupervised Discretization of Continuous Features. *Machine Learning Proceedings 1995*, Elsevier, 1995, p. 194–202. DOI:10.1016/B978-1-55860-377-6.50032-3.

[160] Shahzad A, Mebarki N. Learning Dispatching Rules for Scheduling: A Synergistic View Comprising Decision Trees, Tabu Search and Simulation, *Computers*, 2016, 5, 3. DOI:10.3390/computers5010003.

[161] Blajdo P, Grzymala-Busse JW, Hippe ZS, Knap M, Mroczek T, Piatek L. A Comparison of Six Approaches to Discretization—A Rough Set Perspective. *Rough Sets and Knowledge Technology*, Berlin, Heidelberg, Springer Berlin Heidelberg, 2008, p. 31–8. DOI:10.1007/978-3-540-79721-0_10.

[162] Luque A, Carrasco A, Martín A, de las Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix, *Pattern Recognit*, 2019, 91, 216–31. DOI:10.1016/j.patcog.2019.02.023.

[163] Tharwat A. Classification assessment methods, *Applied Computing and Informatics*, 2021, 17, 168–92. DOI:10.1016/j.aci.2018.08.003.

[164] Novaković JD, Veljović A, Ilić SS, Papić Ž, Tomović M. Evaluation of classification models in machine learning, *Theory and Applications of Mathematics & Computer Science*, 2017, 7, 39.

[165] Vujovic ŽĐ. Classification Model Evaluation Metrics, *International Journal of Advanced Computer Science and Applications*, 2021, 12. DOI:10.14569/IJACSA.2021.0120670.

[166] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, *BMC Genomics*, 2020, 21, 6. DOI:10.1186/s12864-019-6413-7.

[167] Microsoft. Annual Report 2019, Washington, 2019.

[168] Karwińska A. Program nauczania informatyki w klasach IV–VIII szkoły podstawowej, Warszawa, 2019.

[169] Microsoft Corporation. Getting started with VBA in Office 2022. https://learn.microsoft.com/en-us/office/vba/library-reference/concepts/getting-started-with-vba-in-office (dostęp 09/13/2023).

[170] What is the difference between “FMCG” and “CPG?,” *Nielsen Consumer LLC*, 2022. https://nielseniq.com/global/en/insights/analysis/2022/what-is-the-difference-between-fmcg-and-cpg/ (dostęp 04/23/2023).

[171] Pytlarczyk E, Mrówczyński K, Kowalski P, Rykaczewski G. Barometr sektorowy. Podsumowanie 2022 roku w branżach polskiej gospodarki i przewidywania na rok 2023, 2023.

[172] Pieters L, Handrinos N, Cook J, Fenech C, Upadhyaya J. 2023 consumer products industry outlook, 2023.

[173] Brzeziński M. Organizacja produkcji w przedsiębiorstwie 2013.

[174] Mahmoud MA. Classification of production systems, *University of Technology: Baghdad, Iraq*, 2014.

[175] Tayyab M, Habib MS, Jajja MSS, Sarkar B. Economic assessment of a serial production system with random imperfection and shortages: A step towards sustainability, *Comput Ind Eng*, 2022, 171, 108398. DOI:10.1016/j.cie.2022.108398.

[176] Chakravorty SS, Brian Atwater J. A comparative study of line design approaches for serial production systems, *International Journal of Operations & Production Management*, 1996, 16, 91–108. DOI:10.1108/01443579610119117.

[177] Neumann WP, Winkel J, Medbo L, Magneberg R, Mathiassen SE. Production system design elements influencing productivity and ergonomics, *International Journal of Operations & Production Management*, 2006, 26, 904–23. DOI:10.1108/01443570610678666.

[178] Jamal AMM, Sarker BR, Mondal S. Optimal manufacturing batch size with rework process at a single-stage production system, *Comput Ind Eng*, 2004, 47, 77–89. DOI:10.1016/j.cie.2004.03.001.

[179] Buscher U, Lindner G. Optimizing a production system with rework and equal sized batch shipments, *Comput Oper Res*, 2007, 34, 515–35. DOI:10.1016/j.cor.2005.03.011.

[180] Flapper SDP, Fransoo JC, Broekmeulen RACM, Inderfurth K. Planning and control of rework in the process industries: A review, *Production Planning & Control*, 2002, 13, 26–34. DOI:10.1080/09537280110061548.

[181] Liu N, Kim Y, Hwang H. An optimal operating policy for the production system with rework, *Comput Ind Eng*, 2009, 56, 874–87. DOI:10.1016/j.cie.2008.09.013.

[182] WMUROWANIE KAMIENIA WĘGIELNEGO W ŚRODZIE ŚLĄSKIEJ 2021. https://www.pepsicopoland.com/Aktualnosci/Historie/nowy-zak%C5%82ad-pepsico-w-%C5%9Brodzie-%C5%9Bl%C4%85skiej (dostęp 04/24/2023).

[183] Fabryka GSK w Poznaniu jest wiodącym, globalnym producentem tabletek, kapsułek, kremów i maści. n.d. https://pl.gsk.com/pl-pl/kariera/praca-na-produkcji/ (dostęp 04/24/2023).

[184] Goshime Y, Kitaw D, Jilcha K. Lean manufacturing as a vehicle for improving productivity and customer satisfaction, *International Journal of Lean Six Sigma*, 2019, 10, 691–714. DOI:10.1108/IJLSS-06-2017-0063.

[185] Kerdlap P, Low JSC, Ramakrishna S. Zero waste manufacturing: A framework and review of technology, research, and implementation barriers for enabling a circular economy transition in Singapore, *Resour Conserv Recycl*, 2019, 151, 104438. DOI:10.1016/j.resconrec.2019.104438.

[186] Iqbal MW, Kang Y, Jeon HW. Zero waste strategy for green supply chain management with minimization of energy consumption, *J Clean Prod*, 2020, 245, 118827. DOI:10.1016/j.jclepro.2019.118827.

[187] Singh S, Ramakrishna S, Gupta MK. Towards zero waste manufacturing: A multidisciplinary review, *J Clean Prod*, 2017, 168, 1230–43. DOI:10.1016/j.jclepro.2017.09.108.

[188] Awogbemi O, Kallon DV Von, Bello KA. Resource Recycling with the Aim of Achieving Zero-Waste Manufacturing, *Sustainability*, 2022, 14, 4503. DOI:10.3390/su14084503.

[189] International Organization for Standardization. ISO 9000:2015 Quality management systems — Fundamentals and vocabulary 2015.

[190] Hadian SM, Farughi H, Rasay H. Joint planning of maintenance, buffer stock and quality control for unreliable, imperfect manufacturing systems, *Comput Ind Eng*, 2021, 157, 107304. DOI:10.1016/j.cie.2021.107304.

[191] Alcaraz JLG, Sánchez-Ramírez C. Techniques, Tools and Methodologies Applied to Quality Assurance in Manufacturing, Cham, Springer International Publishing, 2021. DOI:10.1007/978-3-030-69314-5.

[192] Ahmad MM, Dhafr N. Establishing and improving manufacturing performance measures, *Robot Comput Integr Manuf*, 2002, 18, 171–6. DOI:10.1016/S0736-5845(02)00007-8.

[193] Radej B, Drnovsek J, Beges G. An overview and evaluation of quality-improvement methods from the manufacturing and supply-chain perspective, *Advances in Production Engineering & Management*, 2017, 12, 388–400. DOI:10.14743/apem2017.4.266.

[194] Wu Z, Liu W, Nie W. Literature review and prospect of the development and application of FMEA in manufacturing industry, *The International Journal of Advanced Manufacturing Technology*, 2021, 112, 1409–36. DOI:10.1007/s00170-020-06425-0.

[195] S. Parsana T, T. Patel M. A Case Study: A Process FMEA Tool to Enhance Quality and Efficiency of Manufacturing Industry, *Bonfring International Journal of Industrial Engineering and Management Science*, 2014, 4, 145–52. DOI:10.9756/BIJIEMS.10350.

[196] Mikulak RJ, McDermott R, Beauregard M. The Basics of FMEA, Taylor & Francis, 2017.

[197] Onodera K. Effective techniques of FMEA at each life-cycle stage. *Annual Reliability and Maintainability Symposium*, IEEE, 1997, p. 50–6. DOI:10.1109/RAMS.1997.571664.

[198] Veena TR, Prabhushankar GV. A literature review on lean, Six Sigma and ISO 9001:2015 in manufacturing industry to improve process performance, *International Journal of Business and Systems Research*, 2019, 13, 162. DOI:10.1504/IJBSR.2019.098652.

[199] Tomic B, Spasojevic Brkic VK. Customer satisfaction and ISO 9001 improvement requirements in the supply chain, *The TQM Journal*, 2019, 31, 222–38. DOI:10.1108/TQM-07-2017-0072.

[200] Brandl FJ, Roider N, Hehl M, Reinhart G. Selecting practices in complex technical planning projects: A pathway for tailoring agile project management into the manufacturing industry, *CIRP J Manuf Sci Technol*, 2021, 33, 293–305. DOI:10.1016/j.cirpj.2021.03.017.

[201] Pettus ML, Kor YY, Mahoney JT. A theory of change in turbulent environments: the sequencing of dynamic capabilities following industry deregulation, *International Journal of Strategic Change Management*, 2009, 1, 186. DOI:10.1504/IJSCM.2009.024509.

[202] Huang GQ, Mak KL. Current practices of engineering change management in UK manufacturing industries, *International Journal of Operations & Production Management*, 1999, 19, 21–37. DOI:10.1108/01443579910244205.

[203] Baskoro G. Designing a Master Program to Cope with the New and Next Normal (VUCA World, Industry 4.0, and Covid 19): a case study, *IPTEK Journal of Proceedings Series*, 2021, 0, 54. DOI:10.12962/j23546026.y2020i3.11078.

[204] Setyanto Putro, Rianto Rianto, Bima Haria Wibisana. MAKING BUSINESS POLICIES AND STRATEGIES IN THE VUCA ERA WITH TECHNOLOGY DEVELOPMENT: A LITERATURE REVIEW, *International Journal of Innovative Technologies in Social Science*, 2022. DOI:10.31435/rsglobal_ijitss/30032022/7796.

[205] Isniah S, Hardi Purba H, Debora F. Plan do check action (PDCA) method: literature review and research issues, *Jurnal Sistem Dan Manajemen Industri*, 2020, 4, 72–81. DOI:10.30656/jsmi.v4i1.2186.

[206] Chong JY, A. Perumal P. Conceptual Framework for Lean Manufacturing Implementation in SMEs with PDCA Approach, 2020, p. 410–8. DOI:10.1007/978-981-13-9539-0_40.

[207] Chiarini A, Cherrafi A. Integrating ISO 9001 and Industry 4.0. An implementation guideline and PDCA model for manufacturing sector, *Total Quality Management & Business Excellence*, 2023, 1–26. DOI:10.1080/14783363.2023.2192916.

[208] Darmawan H, Hasibuan S, Hardi Purba H. Application of Kaizen Concept with 8 Steps PDCA to Reduce in Line Defect at Pasting Process: A Case Study in Automotive Battery, *International Journal of Advances in Scientific Research and Engineering*, 2018, 4, 97–107. DOI:10.31695/IJASRE.2018.32800.

[209] Realyvásquez Vargas A, García Alcaraz JL, Satapathy S, Díaz-Reza JR. Plan-Do-Check-Act Cycle (PDCA) and Auxiliary Tools for Troubleshooting Manufacturing Processes, 2023, p. 1–22. DOI:10.1007/978-3-031-26805-2_1.

[210] Lodgaard E, Gamme I, Aasland KE. Success Factors for PDCA as Continuous Improvement Method in Product Development, 2013, p. 645–52. DOI:10.1007/978-3-642-40352-1_81.

[211] Realyvásquez-Vargas A, Arredondo-Soto K, Carrillo-Gutiérrez T, Ravelo G. Applying the Plan-Do-Check-Act (PDCA) Cycle to Reduce the Defects in the Manufacturing Industry. A Case Study, *Applied Sciences*, 2018, 8, 2181. DOI:10.3390/app8112181.

[212] Du GEJ. Construction Cost Control Based on PDCA Cycle, *International Journal of Simulation: Systems, Science & Technology*, 2016. DOI:10.5013/IJSSST.a.17.37.19.

[213] Bochkovskyi A. Improvement of risk management principles in occupational health and safety, *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu*, 2020, 94–104. DOI:10.33271/nvngu/2020-4/094.

[214] Manavizadeh N, Azizi Javan E. A new approach in joint optimization of maintenance planning, process quality and production scheduling, *International Journal of Research in Industrial Engineering*, 2014, 3, 24–32.

[215] Rahim MA, Ben-Daya M. Joint determination of production quantity, inspection schedule, and quality control for an imperfect process with deteriorating products, *Journal of the Operational Research Society*, 2001, 52, 1370–8. DOI:10.1057/palgrave.jors.2601238.

[216] Dyrektywa Rady 76/211/EWG w sprawie zbliżania ustawodawstw Państw Członkowskich odnoszących się do paczkowania według masy lub objętości niektórych produktów w opakowaniach jednostkowych, Rada Unii Europejskiej, 1976.

[217] Obwieszczenie Marszałka Sejmu Rzeczypospolitej Polskiej z dnia 30 września 2022 r. w sprawie ogłoszenia jednolitego tekstu ustawy o towarach paczkowanych, Marszałek Sejmu RP, 2022.

[218] Obwieszczenie Marszałka Sejmu Rzeczypospolitej Polskiej z dnia 2 grudnia 2020 r. w sprawie ogłoszenia jednolitego tekstu ustawy o produktach biobójczych, Marszałek Sejmu RP, 2020.

10. Spis tabel

11. Spis rysunków

12. Zestawienie omawianych testów

Tabela 12.1. Zestawienie testów, które były omówione w niniejszej pracy.

Numer testu	Nazwa testu	MCC	ACC	ASR	TP	TN	FP	FN
T974	Maksymalne ACC	0,9758	99,85%	95,64%	7188	229	4	7
T975	Maksymalne MCC	0,9758	99,85%	95,62%	7187	229	4	7
T979	Redukcja analiz kosztochłonnych	0,9758	99,85%	95,84%	7204	229	4	7
T981	Redukcja analiz kosztochłonnych	0,9758	99,85%	95,82%	7202	229	4	7
T983	Redukcja analiz kosztochłonnych	0,9758	99,85%	95,65%	7189	229	4	7
T984	Redukcja analiz kosztochłonnych	0,9758	99,85%	95,62%	7187	229	4	7
T978	Redukcja analiz kosztochłonnych	0,9714	99,83%	95,65%	7188	228	5	8
T980	Redukcja analiz kosztochłonnych	0,9714	99,82%	95,62%	7186	228	5	8
T976	Redukcja analiz kosztochłonnych	0,9693	99,81%	95,84%	7202	228	5	9
T977	Redukcja analiz kosztochłonnych	0,9693	99,81%	95,82%	7200	228	5	9
T726	Analiza pH	0,9527	99,72%	91,97%	6915	208	10	10
T724	Analiza pH	0,9504	99,71%	92,52%	6957	208	10	11
T725	Analiza pH	0,9502	99,71%	92,35%	6945	207	11	10
T751	Analiza gęstości	0,9016	99,43%	94,35%	7087	199	23	19
T750	Analiza gęstości	0,9016	99,43%	94,31%	7084	199	23	19
T404	Analiza gęstości	0,8995	99,41%	94,37%	7088	199	23	20
T394	Analiza gęstości	0,8927	99,37%	94,43%	7090	198	24	22
T836	Analiza wolnego chloru	0,8757	99,26%	94,30%	7073	197	25	29
T642	Analiza suchej pozostałości	0,8749	99,25%	94,52%	7087	199	23	32
T671	Dyskretne - analizy sensoryczne	0,8731	99,25%	94,57%	7094	196	26	29
T639	Analiza suchej pozostałości	0,8729	99,24%	94,50%	7085	199	23	33

Numer testu	Nazwa testu	MCC	ACC	ASR	TP	TN	FP	FN
T894	Zmiana min. liczby próbek ogółem	0,8718	99,20%	95,38%	7141	208	26	33
T825	Analiza nadtlenku wodoru	0,8717	99,23%	93,79%	7032	197	24	32
T550	Analiza wolnego chloru	0,8717	99,24%	94,52%	7088	197	25	31
T238	Zmiana max. liczby próbek ogółem	0,8716	99,23%	95,75%	7180	200	23	34
T620	Analiza suchej pozostałości	0,8697	99,22%	94,57%	7091	197	25	32
T239	Zmiana max. liczby próbek ogółem	0,8688	99,21%	96,05%	7199	202	24	35
T607	Analiza suchej pozostałości	0,8678	99,21%	94,59%	7092	197	25	33
T450	Analiza lepkości	0,8672	99,21%	94,44%	7081	196	25	33
T307	Zmiana min. liczby próbek negatywnych	0,8671	98,83%	62,80%	4625	196	24	33
T484	Analiza nadtlenku wodoru	0,8671	99,21%	94,58%	7092	196	26	32
T221	Ogólny punkt odniesienia	0,8651	99,20%	94,57%	7090	196	26	33
T410	Analiza lepkości	0,8651	99,20%	94,57%	7090	196	26	33
T679	Dyskretne - cechy partii	0,8651	99,20%	94,57%	7090	196	26	33
T924	Zmiana max. liczby próbek ogółem	0,8639	99,46%	26,10%	1982	34	10	1
T543	Analiza wolnego chloru	0,8612	99,17%	94,63%	7093	196	26	35
T680	Dyskretne - cechy partii	0,858	99,16%	94,67%	7097	194	28	34
T677	Dyskretne - cechy partii	0,858	99,16%	94,67%	7097	194	28	34
T675	Dyskretne - cechy partii	0,8494	99,12%	95,08%	7130	190	33	32
T678	Dyskretne - cechy partii	0,8494	99,12%	95,08%	7130	190	33	32
T409	Analiza lepkości	0,8488	99,10%	94,67%	7095	192	30	36
T226	Dyskretne - parametry algorytmu	0,8487	99,10%	94,63%	7092	192	32	34
T667	Dyskretne - analizy sensoryczne	0,8439	99,12%	94,57%	7099	181	41	24
T670	Dyskretne - analizy sensoryczne	0,8427	99,09%	94,57%	7092	186	36	31
T314	Zmiana min. liczby próbek negatywnych	0,8423	94,12%	0,66%	37	11	1	2
T672	Dyskretne - analizy sensoryczne	0,8361	99,06%	94,58%	7095	182	40	29

Numer testu	Nazwa testu	MCC	ACC	ASR	TP	TN	FP	FN
T351	Zmiana min. liczby próbek pozytywnych	0,8346	99,23%	69,90%	5278	109	19	23
T925	Zmiana max. liczby próbek ogółem	0,8337	99,28%	32,39%	2455	43	17	1
T674	Dyskretne - cechy partii	0,8328	99,03%	95,31%	7145	186	37	35
T676	Dyskretne - cechy partii	0,8328	99,03%	95,31%	7145	186	37	35
T369	Analiza gęstości	0,8273	98,95%	94,62%	7081	191	31	46
T669	Dyskretne - analizy sensoryczne	0,8255	99,02%	94,58%	7099	175	47	25
T222	Dyskretne - parametry algorytmu	0,8219	98,94%	94,58%	7623	200	31	53
T668	Dyskretne - analizy sensoryczne	0,8217	99,01%	94,59%	7102	172	50	23
T228	Dyskretne - parametry algorytmu	0,8207	98,94%	94,64%	7629	199	34	50
T257	Analiza pH	0,8189	98,94%	94,68%	7093	183	40	38
T666	Dyskretne - analizy sensoryczne	0,8076	98,95%	94,59%	7108	162	60	17
T224	Dyskretne - parametry algorytmu	0,6302	98,03%	96,05%	7185	128	98	49
T665	Dyskretne - parametry algorytmu	0,6283	98,02%	96,05%	7184	128	98	50
T229	Dyskretne - parametry algorytmu	0,6231	97,99%	96,39%	7757	139	96	66
T227	Dyskretne - parametry algorytmu	0,5902	97,85%	96,39%	7756	129	106	67