Uczenie Maszynowe

Klaudia Dynak, Piotr Lipiński, Mikołaj Słupiński

Projekt

Celem projektu jest rozwiązanie wybranego problemu uczenia maszynowego przy użyciu metod poznanych na wykładzie. Kluczowym elementem projektu jest **kreatywność** (pomysłowość, dociekliwość, własna interpretacja problemu, właściwy dobór metod, itp.). Projekty należy wykonywać pod nadzorem prowadzącego zajęcia, w grupach dwu- lub trzyosobowych lub w **szczególnych przypadkach** samodzielnie. Grupy mogą być mieszane tzn. składać się z osób z różnych grup ćwiczeniowych.

Efekt końcowy

Efekt końcowy wykonania projektu powinien zawierać:

Oprogramowanie (skrypty, programy, narzędzia)

- Zarówno w wersji źródłowej, jak i gotowej do uruchomienia (np. Jupyter Notebook z wykonanymi komórkami)
- Nie są istotne sprawy software engineering, formatowania i komentowania kodu, interfejsu użytkownika, itp.
- Mile widziana jest krótka instrukcja uruchomienia i obsługi oprogramowania, jeśli nie jest to oczywiste (plik README.md)

Oryginalne analizowane dane

- Jeśli są wymagane (np. zbiór danych pobrany z Kaggle, UCI ML Repository, lub własne dane)
- W przypadku dużych zbiorów danych wystarczy link do źródła oraz skrypt pobierający dane

Wyniki przeprowadzonych eksperymentów

- Zarówno wyniki ostateczne (omówione w prezentacji), jak i wyniki pośrednie ilustrujące proces modelowania
- Przeprowadzone eksperymenty muszą być możliwe do powtórzenia (należy zapisywać random_state / ziarno generatora)

Prezentacja końcowa

- Zwięzła, ale wyczerpująca prezentacja wyników (format: slajdy lub notebook z komentarzami)
- **Umiejętność opowiadania o problemie jest równie ważna co umiejętność jego rozwiązania**

Zawartość projektu:

Z projektu powinno się dać wywnioskować:

1. **Opis rozpatrywanego zagadnienia** – kontekst biznesowy lub naukowy problemu
2. **Dokładna definicja problemu ML** – typ zadania (klasyfikacja / regresja / klasteryzacja / szeregi czasowe), zmienna docelowa, cechy wejściowe
3. **Eksploracyjna analiza danych (EDA)** – statystyki opisowe, rozkłady zmiennych, korelacje, brakujące wartości, obserwacje odstające, wnioski z analizy

4. **Opis preprocessingu danych** – obsługa brakujących wartości, kodowanie zmiennych kategorycznych, skalowanie, inżynieria cech
5. **Szczegółowy opis użytych modeli** – jeśli nie są to klasyczne algorytmy omówione na wykładzie
6. **Metodologia ewaluacji** – podział danych (train/test/validation), metryki jakości, walidacja krzyżowa
7. **Szczegółowy opis uzyskanych wyników** – porównanie modeli, analiza błędów, wizualizacje
8. **Wnioski końcowe** – podsumowanie, ograniczenia rozwiązania, perspektywy rozwoju

Powyższe powinny jasno wynikać z prezentacji jak i struktury dostarczonego kodu. Tzn. slajdy nie powinny zawierać dokładnego opisu step-by-step jak wygląda preprocessing, natomiast dobrym zwyczajem jest przygotowanie czytelnego Jupyter Notebooka.

Ocena projektu

Projekt zostanie oceniony w skali od 0 do 40 punktów:

Kryterium	Punkty
Wybór metod i narzędzi	5
Efektywność implementacji	6
Otrzymane wyniki końcowe i cząstkowe	6
Wnikliwość analizy danych	10
Prezentacja	8
Całokształt rozwiązania problemu	5
SUMA	**40**

1. Wybór metod i narzędzi (5 pkt)

- Prawidłowa identyfikacja typu problemu (klasyfikacja vs regresja vs klasteryzacja)
- Dobór odpowiednich algorytmów do charakterystyki danych
- Właściwy wybór metryk ewaluacji (np. accuracy vs F1-score dla niezbalansowanych klas)
- Uzasadnienie wyborów metodologicznych

2. Efektywność implementacji (6 pkt)

- Korzystanie z bibliotek (scikit-learn, pandas, numpy) zamiast implementacji od zera
- Wykorzystanie operacji wektorowych zamiast pętli
- Czytelny i reprodukowalny kod (ustawione random_state, logiczna struktura)
- Unikanie typowych błędów (np. data leakage, skalowanie przed podziałem danych)

3. Otrzymane wyniki (6 pkt)

- Jakość uzyskanych predykcji w kontekście trudności problemu
- Dokumentacja procesu iteracyjnego (baseline → ulepszenia)
- Porównanie różnych podejść i modeli
- Analiza błędów modelu

4. Wnikliwość analizy danych (10 pkt)

- Przeprowadzenie rzetelnej eksploracyjnej analizy danych (EDA)
- Identyfikacja i obsługa brakujących wartości, outlierów, duplikatów
- Zrozumienie rozkładów zmiennych i zależności między nimi
- Wnioski z EDA wpływające na dalsze decyzje modelowania
- Sensowna inżynieria cech (feature engineering)

5. Prezentacja (8 pkt)

- Klarowność przekazu i struktura prezentacji
- Umiejętność wyjaśnienia podjętych decyzji
- Wizualizacje wspierające narrację
- Odpowiedzi na pytania prowadzącego

6. Całokształt rozwiązania (5 pkt)

- Spójność całego pipeline'u ML
- Kreatywność i samodzielność w podejściu do problemu
- Krytyczna ocena własnych wyników i świadomość ograniczeń

Uwagi dotyczące zakresu

Metody w zakresie przedmiotu

Projekt powinien wykorzystywać klasyczne metody uczenia maszynowego omawiane na wykładzie:

- Regresja liniowa i logistyczna
- Drzewa decyzyjne (CART, ID3, C4.5)
- Metody zespołowe (Random Forest, Gradient Boosting, XGBoost, LightGBM)
- Support Vector Machines (SVM)
- Naive Bayes
- k-Nearest Neighbors (k-NN)
- Metody klasteryzacji (k-means, DBSCAN, hierarchiczna)
- Redukcja wymiarowości (PCA, t-SNE)

Metody głębokiego uczenia (Deep Learning)

Metody głębokiego uczenia (sieci neuronowe, CNN, RNN, Transformery, itp.) **mogą być użyte jako dodatkowy element projektu**, ale:

- **Nie będą oceniane** jako główna część projektu
- Nie zastępują wymogu użycia klasycznych metod ML
- Mogą stanowić bonus pokazujący szersze umiejętności studenta

Dziedziny wymagające specjalistycznej wiedzy

Zniechęcamy do wybierania projektów z zakresu:

- **Wizji komputerowej (Computer Vision)** – istnieją od tego specjalistyczne przedmioty
- **Przetwarzania języka naturalnego (NLP)** – istnieją od tego specjalistyczne przedmioty

Projekty z tych dziedzin są dopuszczalne tylko po wcześniejszym uzgodnieniu z prowadzącym.

Terminy

Etap	Termin
Ustalenie tematu projektu	**7 stycznia**
Przedstawienie podejścia i pierwszych wyników	**do 17 stycznia**
Przedstawienie końcowej wersji projektu	**do końca semestru**

Konsultacje, pytania, inne ustalenia z prowadzącym zajęcia – cały czas, osobiście lub emailem.

Proponowane tematy projektów

Po szczegółowe informacje proszę indywidualnie kontaktować się z prowadzącym zajęcia.

Klasyfikacja

- **Predykcja churnu klientów** – przewidywanie, którzy klienci zrezygnują z usług
- **Diagnoza medyczna** – klasyfikacja chorób na podstawie objawów lub wyników badań
- **Detekcja fraudów** – wykrywanie oszustw w transakcjach finansowych
- **Klasyfikacja jakości wina** – predykcja jakości wina na podstawie właściwości chemicznych
- **Predykcja sukcesu kampanii marketingowej** – czy klient odpowie na ofertę?

Regresja

- **Predykcja cen nieruchomości** – szacowanie ceny mieszkania/domu
- **Prognozowanie sprzedaży** – przewidywanie wolumenu sprzedaży produktów
- **Predykcja zużycia energii** – prognozowanie zapotrzebowania na energię elektryczną
- **Szacowanie czasu dostawy** – przewidywanie czasu realizacji zamówienia

Szeregi czasowe

- **Prognozowanie cen akcji/kryptowalut** – analiza trendów i predykcja
- **Prognozowanie pogody** – predykcja temperatury, opadów na podstawie danych historycznych
- **Predykcja obciążenia serwerów** – prognozowanie zapotrzebowania na zasoby IT

Klasteryzacja i analiza nienadzorowana

- **Segmentacja klientów** – grupowanie klientów na podstawie zachowań zakupowych
- **Analiza koszykowa** – odkrywanie wzorców zakupowych
- **Detekcja anomalii** – wykrywanie nietypowych obserwacji w danych

Projekty interdyscyplinarne

- **Predykcja wyników sportowych** – analiza statystyk i przewidywanie rezultatów meczy
- **Analiza danych z sensorów IoT** – klasyfikacja aktywności, predykcja awarii maszyn
- **Predykcja jakości powietrza** – prognozowanie zanieczyszczeń na podstawie danych środowiskowych

Kaggle Competitions

- **Spaceship Titanic** – nowsza wersja problemu Titanic
- **Tabular Playground Series** – seria problemów tabelarycznych o różnym poziomie trudności
- **Home Credit Default Risk** – predykcja ryzyka kredytowego
- **IEEE-CIS Fraud Detection** – wykrywanie oszustw
- **Inne aktywne lub archiwalne konkursy** – po uzgodnieniu z prowadzącym

Własne pomysły

Ciekawe pomysły na projekty są mile widziane! Jeśli masz własny pomysł lub dostęp do interesujących danych (np. z pracy, projektu naukowego, hobby), skontaktuj się z prowadzącym w celu ustalenia szczegółów.

W razie pytań zapraszam do kontaktu emailowego na adres `uczeniemaszynowe@cs.uni.wroc.pl` lub na konsultacje.

FAQ

Będzie uzupełniane na bieżąco