



Program studiów podyplomowych
Data Science w zastosowaniach biznesowych – praktyczne warsztaty

PROGRAM STUDIÓW	
Nazwa studiów w języku polskim:	Data Science w zastosowaniach biznesowych - praktyczne warsztaty
Nazwa studiów w języku angielskim:	Data Science in business applications – practical workshops
Język wykładowy:	polski
Liczba semestrów:	2
Łączna liczba godzin zajęć:	226
w tym liczba godzin zajęć wymagana do ukończenia studiów podyplomowych:	226
Łączna liczba punktów ECTS:	60
w tym liczba punktów ECTS wymagana do ukończenia studiów podyplomowych:	60
Poziom PRK, na którym nadawana jest kwalifikacja cząstkowa po ukończeniu studiów podyplomowych:	6

ZASADY ZALICZENIA

Zaliczenie studiów składa się z następujących elementów.

- Testy sprawdzające:** po każdym kursie uczestnicy rozwiązują test dotyczący zagadnień omawianych na zajęciach. Zaliczenie studiów wymaga uzyskania co najmniej 50% możliwych do uzyskania punktów na każdym teście zaliczeniowym.
- Praca dyplomowa:** pod koniec studiów uczestnicy przygotowują pracę dyplomową na temat uzgodniony z wybranym promotorem. Praca może być przygotowana w formie tradycyjnej lub alternatywnie w postaci **aplikacyjnej** (stworzenie i opisanie zbudowanej webowej aplikacji analitycznej) albo **artykułu naukowego**. Wymogiem zaliczenia studiów jest uzyskanie dwóch pozytywnych recenzji (promotora i drugiego recenzenta).



Program studiów podyplomowych
Data Science w zastosowaniach biznesowych – praktyczne warsztaty

PROGRAM ZAJĘĆ

Lp.	Nazwa zajęć	Liczba godzin	Punkty ECTS
1.	Wprowadzenie do środowiska / Wprowadzenie do Pythona	9	2
2.	Wczytywanie danych z różnych źródeł	9	2
3.	Przygotowanie danych do analiz	9	2
4.	Wizualizacja danych	9	2
5.	Statystyczna analiza danych	18	4
6.	Unsupervised Learning	18	4
7.	Zaawansowane programowanie	18	4
8.	Regresja liniowa i logistyczna	18	4
9.	Indywidualne konsultacje dotyczące przygotowania pracy dyplomowej	5	2
10.	Machine learning 1	18	6
11.	Machine learning 2	18	6
12.	Natural Language Processing	18	6
13.	Deep Learning	18	6
14.	Zajęcia fakultatywne – do wyboru 4 * 9 godzin (np. Raporty i prezentacje w R Markdown, Web scrapping, Statystyka bayesowska, Wprowadzenie do szeregów czasowych, Analiza danych przestrzennych, Explainable Artificial Intelligence, Tworzenie aplikacji webowych shiny)	4 * 9	4 * 2
15.	Indywidualne konsultacje dotyczące przygotowania pracy dyplomowej	5	2



Program studiów podyplomowych
Data Science w zastosowaniach biznesowych – praktyczne warsztaty

SZCZEGÓŁOWY PROGRAM ZAJĘĆ

Wszystkie zajęcia mają charakter praktyczny, dzięki czemu studia przygotowują do podjęcia pracy na stanowiskach związanych z szeroko rozumianą analizą danych, budowaniem i walidacją modeli predykcyjnych, wspomaganiem podejmowania decyzji na podstawie danych.

Program studiów obejmuje **226 godzin zajęć**, przypadających na **12 kursów obowiązkowych** (łącznie 180 godzin + 10 godzin seminarium dyplomowego) oraz **4 kursy specjalistyczne** (fakultety – do wyboru przez słuchaczy).

Poniżej znajduje się krótka charakterystyka poszczególnych kursów.

Wprowadzenie do środowiska R / Wprowadzenie do Pythona (9 godzin)

Celem kursu jest wprowadzenie/uporządkowanie wiedzy dotyczącej podstaw programowania w R / Pythonie. Kurs ma stanowić solidną bazę wiedzy i umiejętności niezbędną do opanowania bardziej skomplikowanych problemów prezentowanych na innych kursach, stąd też główny nacisk podczas zajęć zostanie położony na pokazanie efektywnego sposobu pracy z R / Pythonem oraz ogólnie rozumianej analizy danych.

Wczytywanie danych z różnych źródeł (9 godzin)

Celem kursu jest zapoznanie słuchaczy z metodami importu danych do R/pythona. W pierwszej części zaprezentowane zostaną przykłady importowania danych z popularnych formatów plików (np. txt, csv, xls), w tym z tzw. plików tekstowych płaskich (ang. *flat files*), plików ze stałą szerokością kolumn, czy plików, w których jeden rekord zapisany jest w kilku wierszach. Zaprezentowane zostaną także metody warunkowego wczytywania danych, jedynie rekordów spełniających określone warunki. Szczególny nacisk zostanie położony także na efektywne wczytywanie danych z dużych plików. W dalszej kolejności pokazane zostaną metody wczytywania danych z innych pakietów statystycznych (np. SAS, Stata, SPSS, Matlab), a także z baz danych. Słuchacze poznają także sposoby wczytywania danych zapisanych w plikach XML. Ostatnia część kursu będzie poświęcona omówieniu zapisywania (eksportu) danych do omówionych wcześniej popularnych formatów.



Program studiów podyplomowych
Data Science w zastosowaniach biznesowych – praktyczne warsztaty

Przygotowanie danych do analiz (9 godzin)

Celem kursu jest zapoznanie słuchaczy z metodami przygotowania danych do analiz. Jakość analiz statystycznych zależy od jakości danych, na których są przeprowadzane – trudno spodziewać się wiarygodnych wyników analiz, gdy wykorzystane dane są „śmieciowe”. Właściwie przygotowane dane to co najmniej połowa sukcesu ich analizy. W pierwszej kolejności słuchacze poznają metody sprawdzania struktury danych i typów zmiennych oraz ich konwersji do właściwego typu. Omówione zostanie rozróżnienie między postacią długą i szeroką danych, a także sposoby analizy poprawności i spójności danych. Ważnym zagadnieniem, które zostanie poruszone na zajęciach, będą braki danych, które eliminują cały rekord z analiz statystycznych. Słuchacze poznają najczęstsze metody imputacji braków danych – zastępowania ich w sensowny sposób wartościami niebrakującymi.

Statystyczna analiza danych (18 godzin)

Celem zajęć jest zapoznanie uczestników ze statystyczną analizą danych. Kurs ma charakter podstawowy i średniozaawansowany. W pierwszej części Słuchacze dowiedzą się jak wykonać analizę opisową rozkładu empirycznego – zostaną przedstawione zasady, przykłady i interpretacja statystyk opisowych. W drugiej części Słuchacze zapoznają się z procedurą weryfikacji hipotez badawczych na podstawie wnioskowania statystycznego, którą następnie przećwiczą w różnych wariantach na danych empirycznych, przeprowadzając testy statystyczne parametryczne i nieparametryczne dotyczące średniej, wariancji i wskaźnika struktury dla jednej próby, dwóch prób zależnych, niezależnych, a także dla więcej niż dwóch prób.

Wizualizacja danych (9 godzin)

Celem zajęć jest zapoznanie uczestników z podstawowymi funkcjami graficznymi oraz z nowoczesnymi technikami wizualizacji danych. Omówione zostaną najważniejsze rodzaje wykresów pozwalające analizować rozkład zmiennych ilościowych oraz jakościowych oraz zależności między nimi. Ważną częścią kursu jest także pokazanie wykorzystanie R/pythona do tworzenia jakościowej grafiki na potrzeby prezentacji biznesowych czy artykułów naukowych.



Program studiów podyplomowych
Data Science w zastosowaniach biznesowych – praktyczne warsztaty

Zaawansowane programowanie (18 godzin)

Celem kursu jest nauczenie zaawansowanych metod programowania, tworzenia złożonych programów i oceny ich złożoności czasowej, a także tworzenia własnych funkcji i pakietów. W pierwszej kolejności omówione zostanie tworzenie funkcji z wykorzystaniem przetwarzania warunkowego i iteracyjnego (pętle) oraz podstawy programowania zorientowanego obiektowo, w tym tworzenie własnych metod i funkcji. Tematy związane z tworzeniem własnych funkcji zostaną uzupełnione omówieniem zagadnienia programowania defensywnego. Drugą ważną częścią kursu będzie pokazanie narzędzi służących do analizy własnego kodu, oceny jego złożoności czasowej, a także identyfikacji i obsługi błędów oraz sposobów poprawy efektywności. Omówione zostaną także zagadnienia wektoryzacji kodu.

Unsupervised Learning (18 godzin)

Celem przedmiotu jest zapoznanie słuchaczy z możliwościami algorytmów odkrywania wiedzy z danych (Knowledge Discovery in Databases, KDD) wykorzystujących uczenie bez nadzoru (ang. *unsupervised learning*). Przedstawione zostaną główne algorytmy reguł asocjacyjnych (Apriori, Eclat) oraz poszukiwania grup danych przez klastrowanie i możliwość ich wykorzystania w analityce biznesowej. Ich zastosowanie dotyczy głównie danych transakcyjnych (tzw. market basket analysis) i poszukiwania reguł najczęściej występujących wspólnie produktów w koszyku. Realizowane będą modele dla danych rzeczywistych, wymagających przekształcenia i wyczyszczenia danych wejściowych. Wśród metod grupowania omówione zostaną hierarchiczna analiza skupień, metoda k-średnich oraz metody wykorzystujące funkcję gęstości lub model.

Analiza regresji liniowej i logistycznej (18 godzin)

Celem kursu jest wprowadzenie do modelowania ekonometrycznego przez omówienie tradycyjnych modeli parametrycznych - regresji liniowej oraz regresji logistycznej (zarówno binarnej, jak i wielomianowej). Modele te są zwykle punktem odniesienia dla bardziej zaawansowanych algorytmów predykcyjnych. Omówione zostaną podstawy teoretyczne modeli, zagadnienia związane z oceną ich jakości, podstawy podziału na próbę uczącą i testową, a także interpretacji i podstawowej diagnostyki.



Program studiów podyplomowych
Data Science w zastosowaniach biznesowych – praktyczne warsztaty

Machine Learning 1 (18 godzin)

Celem kursu jest wprowadzenie do zagadnień uczenia maszynowego w kontekście klasyfikacji, czyli modelowania zjawisk o charakterze jakościowym i regresji, służącej do modelowania zjawisk o charakterze ilościowym. Klasyfikacja może mieć charakter binarny (dwuwartościowa zmienna zależna) albo wielowartościowy. Kurs stanowi rozwinięcie podstawowych zagadnień uporządkowanych na kursie z analizy regresji liniowej i logistycznej. Szczegółowo omówione zostaną kwestie związane z oceną jakości modelu, podziałem na próbę uczącą i testową, a także walidacją krzyżową modeli. Wśród metod modelowania omówiona zostanie metoda k-najbliższych sąsiadów z zastosowaniem zarówno do problemów klasyfikacji, jak i regresji, maszyna i regresja wektorów nośnych (ang. *support vector machine/ support vector regression*), a także regresja grzbietowa (ang. *ridge regression*) oraz metoda LASSO. Ważną częścią kursu będzie również omówienie kwestii doboru zmiennych do modelu oraz bilansowania próby.

Machine learning 2 (18 godzin)

Kurs jest kontynuacją zajęć Machine Learning 1. Celem zajęć jest omówienie kolejnych metod uczenia maszynowego wykorzystywanych w zagadnieniach klasyfikacyjnych i regresyjnych. W pierwszej kolejności omówione zostaną drzewa decyzyjne i drzewa regresyjne, a następnie lasy losowe. Ważną częścią kursu będą zagadnienia związane z łączeniem/uśrednianiem i "wzmacnianiem" (ang. *bagging and boosting*) modeli, w tym w szczególności zaprezentowany zostanie algorytm eXtreme Gradient Boosting oraz CatBoost. Dodatkowo omówione zostaną także podstawy sztucznych sieci neuronowych. Ostatnia część kursu zostanie poświęcona kwestii strojenia (ang. *tuning*) parametrów modeli.

Natural Language Processing (18 godzin)

Celem kursu jest zapoznanie uczestników z technikami pozwalającymi na analizę danych nieustrukturyzowanych (tekstowych) i przetwarzanie języka naturalnego (ang. *Natural Language Processing, NLP*). Omówione zostanie czyszczenie danych tekstowych, także z wykorzystaniem wyrażeń regularnych oraz podstawowa reprezentacja tekstu, a także wektorowe reprezentacje słów (w2v/glove). W trakcie zajęć zbudowany zostanie przykładowy model NLP, omówione będzie także modelowanie tematów (lda), interpretowanie modeli (lime), a także (opcjonalnie) wizualizacja wielowymiarowych danych tekstowych (t-sne). Omówione zostaną także trendy i najnowsze odkrycia w nlp: np. model BERT.



Program studiów podyplomowych
Data Science w zastosowaniach biznesowych – praktyczne warsztaty

Deep Learning (18 godzin)

Celem kursu jest zapoznanie uczestników z różnymi typami sieci neuronowych i ich zastosowaniami. Podczas zajęć wykorzystywane będą narzędzia Keras i TensorFlow. Słuchacze poznają podstawowe definicje i ideę Deep Learningu, jego wady i zalety. Omówione zostaną najważniejsze funkcje aktywacji (sigmoid, ReLU, tanh, itp.) oraz optymalizacja wag metodą stochastic gradient descent, a także bardziej zaawansowanymi metodami (ADAM, RMSProp, AdaDelta,...). Wspomniana zostanie klasyczna regularyzacja L1/L2 i dropout. Następnie zostaną zaprezentowane metody budowania modeli w Keras oraz optymalizacja hiperparametrów z pakietu tf.nn. Przedstawione zostaną sieci konwolucyjne (warstwy konwolucyjna, pooling i głęboka z przykładami) i ich budowa w Keras wraz ze strojeniem parametrów i omówieniem przykładowych architektur (VGG, ResNet), wizualizacją filtrów i map aktywacji. Ostatnim etapem będzie omówienie sieci rekurencyjnych (LSTM, GRU) i ich zastosowań w Keras.



Program studiów podyplomowych
Data Science w zastosowaniach biznesowych – praktyczne warsztaty

Przykładowe zajęcia fakultatywne

Fakultet 1 – Raporty i prezentacje w R Markdown (9 godzin)

Celem fakultetu jest zaprezentowanie Słuchaczom modułu R Markdown – służącego do generowania automatycznych raportów na podstawie kodu źródłowego sporządzonego w R. R Markdown umożliwia pisanie roboczych raportów, artykułów naukowych, prostych stron internetowych, dokumentów MS Word a prezentacji i dokumentów PDF z wykorzystaniem LaTeX z poziomu R Studio. Podczas zajęć Słuchacze będą zapoznawać się z kolejnymi zagadnieniami, a następnie z pomocą prowadzącego wykonywać przykładowe ćwiczenia.

Fakultet 2 – Web scraping (9 godzin)

Celem fakultetu jest zaprezentowanie Słuchaczom technik web scrapingu – pozyskiwania informacji do analiz bezpośrednio ze stron i aplikacji internetowych. Web scraping pozwala znakomicie skrócić czas poświęcany na pozyskanie i/lub aktualizację danych internetowych. Podczas zajęć Słuchacze będą zapoznawać się z kolejnymi zagadnieniami, a następnie z pomocą prowadzącego wykonywać przykładowe ćwiczenia.

Fakultet 3 – Statystyka bayesowska (9 godzin)

Celem fakultetu jest wprowadzenie do statystyki bayesowskiej poprzez omówienie kilku popularnych modeli statystycznych. Analiza bayesowska proponuje odmienne od klasycznej statystyki podejście do wnioskowania statystycznego – jest szeroko wykorzystywana w naukach przyrodniczych i społecznych.

Fakultet 4 – Wprowadzenie do szeregów czasowych (9 godzin)

Celem fakultetu jest wprowadzenie Słuchaczy w analizę szeregów czasowych. Jest to typ danych występujący masowo m. in. w finansach, biznesie i makroekonomii. Znajomość technik analizy i estymacji danych czasowych jest zatem kluczowa w wielu zadaniach analitycznych.



Program studiów podyplomowych
Data Science w zastosowaniach biznesowych – praktyczne warsztaty

Fakultet 5 – Analizy przestrzenne w R (9 godzin)

Celem fakultetu jest poznanie przez Słuchaczy technik analizy danych o charakterze przestrzennym, tj. przypisanych do określonej lokalizacji w przestrzeni geograficznej. Słuchacze poznają sposoby wizualizacji danych przestrzennych (mapy dla danych punktowych i regionalnych) oraz takie operacje jak liczenie odległości między punktami czy wskazanie punktów w zadanej odległości czy obszarze. Wykorzystane zostaną pakiety spdep, sp, rgdal, ggmap, maptools, smarterPoland.

Fakultet 6 – Tworzenie aplikacji webowych (9 godzin)

Celem kursu jest zapoznanie Słuchaczy z tworzeniem interaktywnych aplikacji w projekcie R. Aplikacje Shiny są w całości przygotowywane w pakiecie R, bazując na jego algorytmach i pakietach. Dzięki wykorzystaniu interfejsu graficznego (widgety) możliwe jest tworzenie aplikacji z interfejsem użytkownika dla odbiorców niezainteresowanych ingerencją w kod R. Jednocześnie jest to znakomity sposób na skrócenie wykonywania powtarzalnych analiz oraz prezentacji wyników w formie interaktywnej, z opcją samodzielnego doboru parametrów przez użytkownika.

Fakultet 7 – Wytłumaczalne uczenie maszynowe (9 godzin)

Celem kursu jest zaprezentowanie narzędzi wytłumaczalnego uczenia maszynowego, które pozwalają zinterpretować modele mające charakter "czarnej skrzynki". W pierwszej kolejności omówiona zostanie interpretacja globalna modelu: ocena jakości modelu (analiza reszt), ocena ważności zmiennych (lokalne metody oceny ważności zmiennych, permutacyjne metody oceny ważności zmiennych), analizy scenariuszowe typu „co jeśli” (Partial Dependence Plot, Accumulated Local Effects). W drugiej części zajęć wykorzystane zostaną metody lokalnej interpretacji modelu: Local Surrogate (LIME), Shapley Values, SHAP (SHapley Additive exPlanations), BreakDown.



Program studiów podyplomowych
Data Science w zastosowaniach biznesowych – praktyczne warsztaty

EFEKTY UCZENIA SIĘ

	WIEDZA	
W1	Absolwent studiów ma zaawansowany stopień znajomości środowiska programistycznego R lub python, w tym zna zasady budowania skomplikowanych algorytmów i optymalizacji (profilowania) kodu.	P6S_WG
W2	Absolwent studiów ma zaawansowany stopień znajomości działania algorytmów wykorzystywanych w uczeniu maszynowym (ang. <i>machine learning</i>), uczeniu głębokim (ang. <i>deep learning</i>) czy przetwarzaniu języka naturalnego (ang. <i>natural language processing</i>).	P6S_WG
W3	Absolwent studiów ma zaawansowany stopień znajomości teorii statystycznej i matematycznej potrzebnej do prawidłowego wnioskowania statystycznego w metodach ilościowych.	P6S_WG
W4	Absolwent studiów zna terminologię używaną w Data Science oraz jej zastosowanie w dyscyplinach pokrewnych na poziomie rozszerzonym.	P6S_WG
W5	Absolwent studiów zna zasady i techniki poprawnej i efektywnej wizualizacji danych.	P6S_WG
W6	Absolwent studiów rozumie korzyści z zastosowania metod statystycznych w aplikacjach biznesowych.	P6S_WK
W7	Absolwent studiów zna nowoczesne metody przetwarzania danych	P6S_WK
W8	Absolwent studiów zna specyfikę gromadzenia i wykorzystania danych i algorytmów statystycznych z poszanowaniem praw autorskich, własności przemysłowej i ochrony danych osobowych.	P6S_WK
W9	Absolwent studiów zna problemy, procesy, trendy i wyzwania we współczesnej analityce biznesowej.	P6S_WK
W10	Absolwent studiów zna publikacje, czasopisma i inne źródła wiedzy o zastosowaniu Data Science w biznesie.	P6S_WK
	UMIEJĘTNOŚCI	
U1	Absolwent studiów potrafi identyfikować i formułować na bazie obserwowanych zjawisk problemy badawcze możliwe do weryfikacji na podstawie dostępnych danych ilościowych.	P6S_UW
U2	Absolwent studiów potrafi dobrać i zastosować do rozwiązania danego	P6S_UW



Program studiów podyplomowych
Data Science w zastosowaniach biznesowych – praktyczne warsztaty

	problemu badawczego odpowiednie narzędzie statystyczne.	
U3	Absolwent studiów potrafi poprawnie wykorzystywać algorytmy uczenia maszynowego, uczenia głębokiego, przetwarzanie języka naturalnego oraz modelowania ekonometrycznego.	P6S_UW
U4	Absolwent studiów potrafi budować zaawansowane kody i procedury w języku R lub python do analizy danych automatyzujące pracę.	P6S_UW
U5	Absolwent studiów potrafi opracowywać i komunikować wyniki swoich analiz statystycznych, a także potrafi opracowywać raporty i prezentacje prowadzonych prac badawczych, także w języku angielskim.	P6S_UK
U6	Absolwent studiów potrafi współdziałać i pracować w grupach i projektach analitycznych, przyjmując w nich różne role.	P6S_UG
U7	Absolwent studiów potrafi planować pracę analityczną, odpowiednio określić priorytety służące realizacji określonego przez siebie lub innych zadania.	P6S_UG
U8	Absolwent studiów potrafi pracować samodzielnie oraz zespołowo w instytucjach wykorzystujących metody analizy i przetwarzania danych.	P6S_UG
U9	Absolwent studiów potrafi łączyć wiedzę teoretyczną z kompleksowym podejściem do danych – od pozyskania, przetwarzania, analizowania z użyciem metod wnioskowania statystycznego, ekonometrii, wizualizacji, po raportowanie i prezentację danych.	P6S_UG
U10	Absolwent studiów potrafi zdefiniować i poszerzać wiedzę i umiejętności w obszarze Data Science, a także potrafi ukierunkowywać i planować rozwój innych osób w zakresie Data Science.	P6S_UU
	KOMPETENCJE SPOŁECZNE	
K1	Absolwent studiów podyplomowych na podstawie przedstawionych interpretacji uzyskanych wyników potrafi być krytyczny w stosunku do przedstawionych modeli, prawidłowo identyfikuje i rozstrzyga dylematy wykorzystaniem tych metod w prowadzeniu własnej firmy lub pracy zawodowej.	P6S_KK
K2	Absolwent studiów podyplomowych potrafi dokonać merytorycznej oceny i weryfikacji poprawności stosowanych metod w aplikacjach biznesowych w środowisku R lub python.	P6S_KK
K3	Absolwent studiów podyplomowych jest przygotowany do podjęcia pracy	P6S_KK



Program studiów podyplomowych
Data Science w zastosowaniach biznesowych – praktyczne warsztaty

	w charakterze wysoko wykwalifikowanego analityka w różnych branżach (m.in. finanse, bankowość, ubezpieczenia, badania rynkowe, energetyka, sektor publiczny).	
K4	Absolwent studiów jest gotów do inicjowania projektów analitycznych wspomagających działania na rzecz interesu publicznego, a także działalność przedsiębiorczą.	P6S_KO
K5	Absolwent studiów jest gotów do profesjonalnego i odpowiedzialnego wypełniania obowiązków analityka danych (ang. <i>data scientist</i>).	P6S_KR
K6	Absolwent studiów jest gotów do przestrzegania i rozwijania zasad etyki zawodowej (rzetelności badawczej) i popierania działania na rzecz przestrzegania wysokich standardów badań statystycznych z poszanowaniem zasad prawnych, m. in. praw własności intelektualnej, ochrony danych osobowych i tajemnicy statystycznej.	P6S_KR