



UNIWERSYTET  
WARSZAWSKI



WYDZIAŁ NAUK  
EKONOMICZNYCH

## PROGRAM STUDIÓW

Studia składają się z 2 semestrów. Semestr 1 ma charakter wprowadzający, średniozaawansowany, natomiast semestr 2 w całości poświęcony jest zaawansowanym zajęciom specjalistycznym.

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, wspomaganie podejmowania decyzji na podstawie danych.

**Praca dyplomowa** składana na zakończenie studiów może być przygotowana w formie **tradycyjnej** lub alternatywnie w postaci **aplikacyjnej** (stworzenie i opisanie zbudowanej webowej aplikacji analitycznej) albo **artykułu naukowego**.

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

### Semestr I (zimowy)

- Wprowadzenie do środowiska R – 9 godzin
- Wczytywanie danych do R – 9 godzin
- Przygotowanie danych do analiz – 9 godzin
- Wizualizacja danych – 9 godzin
- Statystyczna analiza danych – 18 godzin
- Unsupervised Learning – 18 godzin
- Zaawansowane programowanie w R – 18 godzin
- Regresja liniowa i logistyczna – 18 godzin
- Indywidualne konsultacje dotyczące przygotowania pracy dyplomowej - 5 godzin

### Semestr II (letni)

- Machine learning 1 – 18 godzin
- Machine learning 2 – 18 godzin
- Natural Language Processing – 18 godzin
- Deep Learning w R – 18 godzin
- 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)
- Indywidualne konsultacje dotyczące przygotowania pracy dyplomowej - 5 godzin



UNIwersytet  
Warszawski



WYDZIAŁ NAUK  
EKONOMICZNYCH

## ZAJĘCIA OBOWIĄZKOWE

### Wprowadzenie do R (9 godzin)

Celem kursu jest prezentacja podstawowych możliwości programu R oraz interfejsu RStudio. 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 tym programem oraz ogólnie rozumianej analizy danych. Kurs ten będzie prowadzony wg zasady 20/80 tj. 20 procent zasobów programistycznych jest używanych w 80% procentach przypadków. Kurs ten zakłada zerową wiedzę z zakresu posługiwania się programem R. Studenci opanują posługiwanie się pakietami programu R.

### Wczytywanie danych do R (9 godzin)

Celem kursu jest zapoznanie słuchaczy z metodami importu danych do R. W pierwszej części zaprezentowane zostaną przykłady importowania danych do R 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, w związku z czym omówione zostaną podstawy wykorzystania języka SQL w R. Słuchacze poznają także sposoby wczytywania danych zapisanych w plikach XML. Ostatnia część kursu będzie poświęcona omówieniu zapisywania (eksportu) danych z R do omówionych wcześniej popularnych formatów.

### 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 jakości danych oraz ich czyszczenia z wykorzystaniem tzw. wyrażeń regularnych, nauczą się sprawdzać czy dane zapisane w tabeli odpowiadają przyjętemu wzorcowi (np. właściwy kod pocztowy, data, adres, nr PESEL, itp.). Kolejnym 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.



UNIwersytet  
Warszawski



WYDZIAŁ NAUK  
EKONOMICZNYCH

### **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 wyłożone 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.

### **Wizualizacja danych (9 godzin)**

Celem zajęć jest zapoznanie uczestników z podstawowymi funkcjami graficznymi w R oraz z nowoczesnymi technikami wizualizacji danych w pakiecie ggplot2. Zajęcia mają charakter podstawowy i średniozaawansowany w ramach kursu wykładana jest zarówno składnia funkcji z pakietu ggplot2, jak i zasady poprawnej i efektywnej wyników analizy statystycznej

### **Zaawansowane programowanie w R (18 godzin)**

Celem kursu jest nauczenie zaawansowanych metod programowania w R, 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 for, while, repeat), koncepcja środowiska i zakresu nazw używanych wewnątrz funkcji oraz podstawy programowania zorientowanego obiektowo w R - klasy systemu S3 i S4, w tym tworzenie własnych metod i funkcji generycznych w obu systemach. 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 zagadnienia wektoryzacji kodu, funkcje z rodziny "apply" oraz wykorzystanie podstaw języka C++ w R (pakiet Rcpp) m.in. do zastąpienia pętli. Zaprezentowane zostaną również narzędzia pozwalające w prosty sposób zbudować własny pakiet.



UNIwersytet  
Warszawski



WYDZIAŁ NAUK  
EKONOMICZNYCH

### **Regresja liniowa i logistyczna (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.

### **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 jako pierwsza omówiona zostanie liniowa i kwadratowa analiza dyskryminacji (jako uzupełnienie metod parametrycznych). W dalszej części kursu przedstawiona zostanie również metoda k-najbliższych sąsiadów z zastosowaniem zarówno do problemów klasyfikacji, jak i regresji, maszyna wektorów nośnych (ang. *support vector machine*), 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.



UNIWERSYTET  
WARSZAWSKI



WYDZIAŁ NAUK  
EKONOMICZNYCH

### **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.

### **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.

### **Deep Learning w R (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 pakieten tfruns. Przedstawione zostaną sieci konwolucyjne (warstwy konwolucyjna, poolingu 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.



UNIwersytet  
Warszawski



WYDZIAŁ NAUK  
EKONOMICZNYCH

## **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 scrapping (9 godzin)**

Celem fakultetu jest zaprezentowanie Słuchaczom technik web scrappingu – pozyskiwania informacji do analiz bezpośrednio ze stron i aplikacji internetowych. Web scrapping 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 z wykorzystaniem pakietu R. 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.



UNIwersytet  
Warszawski



Wydział Nauk  
Ekonomicznych

#### Fakultet 5 – **Analiza danych przestrzennych** (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 – **Algorytmiczne strategie inwestycyjne - wprowadzenie** (9 godzin)

Celem kursu jest omówienie teoretycznych podstaw i pokazanie praktycznych przykładów implementacji podstawowych algorytmicznych strategii inwestycyjnych na danych dziennych oraz śróddziennych. Uczestnicy nauczą się przygotowywać, testować i wdrażać ilościowe strategie inwestycyjne, poznają charakterystyki danych o wysokiej częstotliwości oraz narzędzia wykorzystywane do oceny zyskowności strategii inwestycyjnych. Wszystkie zagadnienia będą ilustrowane praktycznymi przykładami i ćwiczeniami w R.

#### Fakultet 7 – **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 (widżety) 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 8 – **Explainable Artificial Intelligence** (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.